Joint Spatial-Angular Sparse Coding, Compressed Sensing, and Dictionary Learning for Diffusion MRI

by

Evan Schwab

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2018

Abstract

Neuroimaging provides a window into the inner workings of the human brain to diagnose and prevent neurological diseases and understand biological brain function, anatomy, and psychology. Diffusion Magnetic Resonance Imaging (dMRI) is an emerging medical imaging modality used to study the anatomical network of neurons in the brain, which form cohesive bundles, or fiber tracts, that connect various parts of the brain. Since about 73% of the brain is water, measuring the flow, or diffusion of water molecules in the presence of fiber bundles, allows researchers to estimate the orientation of fiber tracts and reconstruct the internal wiring of the brain, *in vivo*.

dMRI signals can be modeled within two domains: the spatial domain consisting of voxels in a brain volume and the diffusion or angular domain, where fiber orientation is estimated in each voxel. Researchers aim to estimate the probability distribution of fiber orientation in every voxel of a brain volume in order to trace paths of fiber tracts from voxel to voxel over the entire brain. Therefore, the traditional framework for dMRI processing and analysis has been from a voxel-wise vantage point with added spatial regularization considered post-hoc. In contrast, we propose a new joint

spatial-angular representation of dMRI data which pairs signals in each voxel with the global spatial environment, jointly. This has the ability to improve many aspects of dMRI processing and analysis and re-envision the core representation of dMRI data from a local perspective to a global one.

In this thesis, we propose three main contributions which take advantage of such joint spatial-angular representations to improve major machine learning tasks applied to dMRI: sparse coding, compressed sensing, and dictionary learning. First, we will show that we can achieve sparser representations of dMRI by utilizing a global spatial-angular dictionary instead of a purely voxel-wise angular dictionary. As dMRI data is very large in size, we provide a number of novel extensions to popular spare coding algorithms that perform efficient optimization on a global-scale by exploiting the separability of our dictionaries over the spatial and angular domains. Next, compressed sensing is used to accelerate signal acquisition based on an underlying sparse representation of the data. We will show that our proposed representation has the potential to push the limits of the current state of scanner acceleration within a new compressed sensing model for dMRI. Finally, sparsity can be further increased by learning dictionaries directly from datasets of interest. Prior dictionary learning for dMRI learn angular dictionaries alone. Our third contribution is to learn spatial-angular dictionaries jointly from dMRI data directly to better represent the global structure. Traditionally, the problem of dictionary learning is non-convex with no guarantees of finding a globally optimal solution. We derive the first theoretical

results of global optimality for this class of dictionary learning problems.

We hope the core foundation of a joint spatial-angular representation will open a new perspective on dMRI with respect to many other processing tasks and analyses. In addition, our contributions are applicable to any general signal types that can benefit from separable dictionaries. We hope the contributions in this thesis may be adopted in the larger signal processing, computer vision, and machine learning communities.

Primary Readers:

Professor René Vidal, Chair (BME/ECE)

Assistant Professor Nicolas Charon (AMS)

Secondary Readers:

Professor Trac Tran (ECE)

Assistant Professor Archana Venkataraman (ECE)

Acknowledgments

The pursuit of my Ph.D. has been one of the most rewarding and fulfilling journeys in my life and for that I have many important people to thank.

I would like to start with the person who has impacted me the most: my advisor Dr. René Vidal. From the beginning of my research career until the very end, René has always upheld the highest of standards. Though admittedly fearful of his level of demand at the start, I quickly grew to appreciate his unmatched style of advising: generously lending hours of his limited time (even throughout the night), giving his full attention to all facets of a problem, zeroing in on every important detail while simultaneously stepping back to survey the big picture, and his ability to clearly articulate and distill complicated topics by making comparisons to many diverse fields. René has pushed me to always be prepared, to study topics with an in-depth level of understanding, and to work the hardest I have ever worked. I thank you René for equipping me with the tools to be successful and bringing out the best in me. I am so lucky to have had you as an advisor and a mentor for the past 6 years and I always thought if I could make it with you, I could make it anywhere.

ACKNOWLEDGMENTS

Next I would like to thank my co-advisor Dr. Nicolas Charon. After seeing his inaugural talk as a new assistant professor at JHU, I knew we had a lot of research interests in common. I am so thankful that Nicolas agreed to advise me on my thesis project since his deep mathematical background and standard for important and impactful work was everything for me to elevate my research. For Nicolas too, no detail was too small to address and countless hours of his time were spent with me on long productive research discussions. Though technically only one year my senior, Nicolas is a prodigy of mathematics and technical research, and I am forever grateful for his never-ending support and guidance throughout the second half of my Ph.D.

I would next like to thank my committee members. Dr. Trac Tran has been a cherished academic support throughout my Ph.D., serving first on my qualifying exam, then on my graduate board oral exam, and finally on my thesis committee. It all began when I took Dr. Tran's amazingly well-taught course on compressed sensing which directly inspired my future research directions and interests. It is an honor to have you on my thesis committee as I present to you my work on compressed sensing. Then, I would like to thank Dr. Archana Venkataraman for serving on my thesis committee. In our brief overlap at JHU, I was lucky to see one of her inspiring talks on brain connectivity and I knew we may share some research interests. Thank you for investing the time to read my thesis and impart your knowledge in this research area.

I would also like to give my deepest thanks to Dr. Michael Miller. I recall sitting

down to meet with Dr. Miller at his big desk when I first joined the Center for Imaging Science at JHU and receiving the warmest, most welcoming and enthusiastic invitation to his center. That personality persists every single day that I see Dr. Miller and it reminds me of the close-knit family he has created at one of the world's leading institutions for medical imaging research. Dr. Miller's far-reaching visions combined with his attention for detailed research, and his devotion the people of CIS, has inspired in me a model of what true leadership looks like. I have been privileged to be a part of your CIS family, Dr. Miller.

I would also like to thank the many colleagues and lab mates that have come and gone over the years who have helped and supported me in and outside the lab: Ertan Cetingül, for being my first mentor during my Ph.D. when I adopted his project on diffusion MRI. As I followed in his footsteps in research and worked with him for two summers as an intern at Siemens in Princeton New Jersey, Ertan was not only a great advisor but a true friend. Bijan Afsari became my trusted post-doc, answering all of my questions (all the time) and preparing me for René's meetings during my first few years in the lab. Though Bijan's mathematical talent and welcoming support inspired in me many ideas, his camaraderie and hilarious sense of humor are what will last in my memory. Finally, my CIS lab mates past and present: Giann Gorospe, Siddharth Mahendran, Linling Tao, Chong You, Manolis Tsakiris, Colin Lea, Flori Yellin, Efi Mavroudi, Connor Lane, Daniel Tward, Kwame Kutten, Neil Hallonquist, Ehsan Jahangiri, Bahman Afsari, Luca Zappella, Benjamin Bejar, Rizwan Chaudhry,

Ehsan Elhamifar, and Roberto Tron. A special acknowledgement goes to Ben Haeffele for his support during the writing of this thesis.

Finally, I'd like to thank my family: my parents Robin and Barry for their loving support throughout my life and my Ph.D. I would not be as motivated, dedicated, educated, and patient without your upbringing. My brothers Nathan and Corey for their love and fun times and my dogs Teddy and Bernie for their stress-relief.

Most of all, I have to thank my wife Amanda for her constant unwavering support during this major life undertaking, dealing with my stresses and doubts, celebrating me for my accomplishments and supporting my career goals. Amanda your love has been my lifeline during this Ph.D. and I cannot tell you enough how much I love and appreciate you.

Thank you.

Dedication

This thesis is dedicated to my dad. It was your Ph.D. and incredible career that inspired me to do well in high school, do well in college, complete my Ph.D. and have a fulfilling career and life just like yours.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The brain is one of the most mysterious and well-studied organs in human history and understanding how it works has captivated and eluded humans for centuries. From psychology, to biology, neuroscience, medicine and engineering, the human brain has been a consistent focal point of various fields of scientific research, each attempting to answer the questions of how the brain works, how can we cure brain diseases, how can we mimic the biology of the brain to build things in our world, and what is consciousness? Early research consisted of studying the brain anatomy of post-mortem human and animal subjects utilizing new tools invented for careful preservation and dissection. With the advent of psycho-analysis, emphasis was placed on understanding the different brain functions such as speech, emotion, and motor-function by invasive surgical experiments and electro-therapy. Then, the 20th century saw a milestone of technological advances in the way of non-invasive neuro-imaging

such as X-ray, PET, CT, and MRI, to name but a few modalities, allowing scientists to image the inside of living human brains. As an analogy to the world of astronomy, neuro-imaging has been to the brain as the telescope has been to the galaxy of the stars. With the ability to visualize brain function and anatomy through non-invasive imaging technologies, advances in brain research have been expanding exponentially.

The turn of the 21st century has now seen another monumental advance in neuro-imaging technology: the invention of a medical imaging modality called diffusion magnetic resonance imaging (dMRI), which has the unique ability to visualize the complex universe of neurons in the brain, intertwined in a vast web of connected electro-chemical pathways. This amazing view into the structural neuroanatomy of the brain allows scientists to map and quantify the circuitry of our minds to better understand how the brain is wired and how the anatomy changes in the face of injury or disease. This technology has been highly beneficial in discovering disease biomarkers, or features of neurological diseases pathology that signify the presence of a disease or predict its onset.

At the same time, another monumental technology called machine learning[1] has taken the world by storm. Machine learning is a class of computational techniques at the intersection of computer science, mathematics and statistics that allow computers to learn to solve problems based on experience without being explicitly programmed. One group of machine learning problems is known as *supervised learning* where a

[1]Other common names include *pattern recognition*, *data mining*, and *artificial intelligence*.

machine is taught to perform a task based on positive and negative examples related to the task. Just as humans learn how to distinguish between horses and zebras based on their characteristics by seeing repeated examples of each animal, a computer too can learn to correctly identify a horse or a zebra in an image based on previous examples of horse and zebra images. For supervised learning the examples used to train the computer have labels (e.g. horse or zebra) which is akin to a teacher supervising a child's learning process by confirming the name of the animal they see. Another type of learning is called *unsupervised learning* which tries to discover structure (e.g. clusters, low-dimensional manifolds, sparse representations) from data without training labels. In this setting we can teach a computer to discover patterns in data in order to make predictions when presented with a new information. In the current digital age, massive amounts of data are becoming exponentially more available, storable, and downloadable with greater ease, and so making sense of "big data" using machine learning is becoming increasingly more vital in all facets of life, like business, commerce, social media, biotechnology, robotics, engineering and medicine.

In the medical domain, machine learning has become a very beneficial tool for radiologists and clinicians to aid in disease diagnosis and treatment recommendations based on medical histories. For neuro-imaging, researchers are using machine learning to identify neuro-anatomical biomarkers in complex dMRI data in order to study Parkinson's and Alzheimer's disease and disorders like traumatic brain injury. In this

thesis, I will apply novel methods from machine learning to dMRI data in order to overcome some of the research challenges outlined in the next section.

1.1 Research Challenges

The relatively young field of dMRI is a burgeoning research domain with a vast number of open challenges. The main goal of dMRI is to measure the flow, or diffusion, of water in the brain. The human brain contains billions of neurons organized in a vast network of anatomical connections. Since water will diffuse preferentially along the directions of structural objects like bundles of neurons, called fibers (see Figure 1.1), by measuring directions of water diffusion, researchers can estimate the local orientation of fibers and use it to reconstruct the entire anatomical network of fiber tracts in the brain, *in vivo*, with a method known as tractography (see Figure 1.2). In order to accurately reconstruct anatomical fiber tracts, a major research challenge is devising accurate models to estimate fiber orientations. Furthermore, the processing of dMRI signals to interpretable biological information is highly dependent on the accuracy and robustness of the underlying diffusion model. Tasks such as de-noising, fiber tract segmentation, registration, and atlas construction are important processing components to standardize the diffusion data for population studies, statistical feature analyses and disease classification via machine learning. Through statistical connectivity analyses and the integration of information from multiple imaging

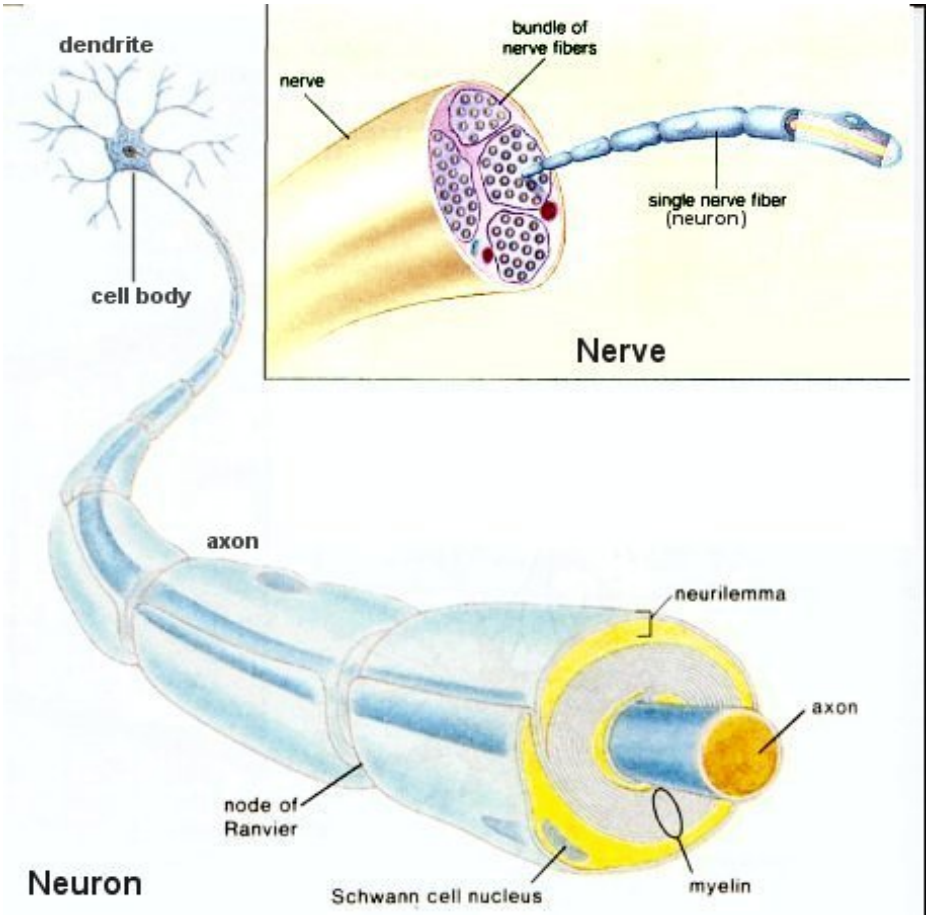


Figure 1.1: Groups of nerve cells form fiber bundles, transmitting electro-chemical signals in a vast anatomical web within the brain. Researchers are able to estimate the orientation and connectivity of the neuroanatomy, *in vivo*, by measuring the restricted water diffusion caused by these fiber bundles.

modalities, dMRI offers a quantitative and qualitative window into the anatomical blueprint of the human mind.

However, before all of these important analyses can take place, dMRI must be proven to be an accessible imaging modality. As it stands at the time of this writing, dMRI is an emerging technology more common in academic research laboratories than in hospitals or clinical arenas. Unlike other frequently used modalities like

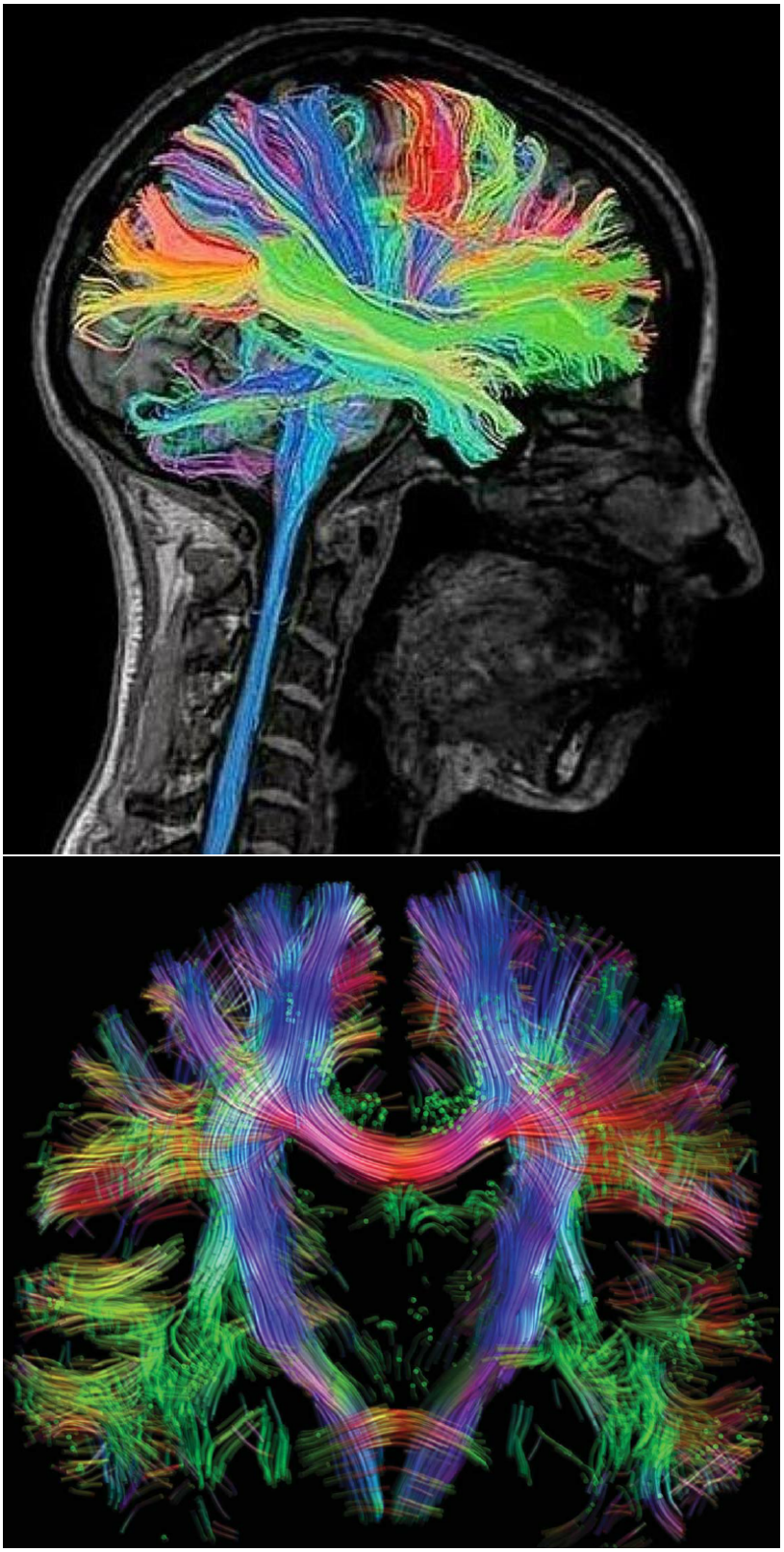


Figure 1.2: Fiber Tractography in the human brain. [Second image adapted from [1].]

MRI and CT, the more involved nature of dMRI results in lengthier scanning times sometimes hundreds of times slower than traditional MRI. Therefore, a major goal is the acceleration of dMRI signal acquisition, which can be achieved by addressing the following three research challenges:

1. ***Sparse Coding:*** *Discovering a representation of dMRI data that is as sparse as possible.*

2. ***Compressed Sensing:*** *Exploiting a sparse representation to reduce the number of measurements needed to recover a full high resolution signal.*

3. ***Dictionary Learning:*** *Learning representations directly from dMRI data to discover sparser codes that are unique to the structure of dMRI.*

Devising novel methods to address these challenges will be the main contributions in this thesis, with the main application of accelerating dMRI to a more clinically feasible level.

There have been a multitude of recent works which have made great gains in the above research challenges, but, they have experienced inherent limitations to the amount of acceleration that can be achieved. A reason for this is that dMRI has traditionally been viewed as a collection of diffusion measurements within each voxel[2] in a brain volume. Accordingly, diffusion modeling, processing, and analysis have been based in a voxel-wise viewpoint. In this thesis, we will show that this

[2]Analogous to the discretization of images into an array of 2D pixels (or *picture elements*), a voxel (or *volume element*) is a 3D pixel for the discretization of volume images.

collection of local representations of dMRI limits the amount of acceleration that can be achieved during acquisition due to the redundancies present between diffusion signals in surrounding spatial neighborhoods. More generally, we will consider a new global framework for diffusion modeling which may impact may other aspects of dMRI processing and have important machine learning applications in the age of big data. In the next section, I outline the core contributions and the organization of this thesis.

1.2 Outline of Thesis Contributions

In this thesis, I propose three main contributions presented in Chapters 3, 4 and 5. The foundational innovation at the core of each contribution is a new global framework for representing dMRI data, that exploits the unique structure of this complex, large-scale neuro-imaging data. In particular, dMRI has two main domains: the spatial domain captures variations in water diffusion across different 3D locations of the MRI volume, and the angular domain captures the main directions of water diffusion at each spatial location. While voxel-wise representations are referred to as purely *angular*, our global representation will be known as jointly *spatial-angular*. The proposed global representation will be used to build three novel machine learning frameworks applied to the joint *spatial-angular* domain of dMRI.

First, in Chapter 2, we will provide background on the main themes of this thesis. In particular, in Chapter 2.1, we will give a mathematical overview of the principles

of dMRI, from acquisition, diffusion modeling and signal reconstruction, to methods for acceleration. Then in Chapter 2.2, we will review the machine learning topics that form the foundations for the main contributions of this thesis: *sparse coding* (Chapter 3), *compressed sensing* (Chapter 4), and *dictionary learning* (Chapter 5) and future directions in *convolutional* methods (Chapter 6). We conclude in Chapter 7.

1.2.1 Spatial-Angular Sparse Coding

For the first major contribution, in Chapter 3, we introduce our proposed joint spatial-angular representation of dMRI and demonstrate its first application in the machine learning area of *sparse coding*. Specifically, the goal of sparse coding is to find a code, or representation of the data, that has a sparse, or very few, number of elements. Since dMRI is very large and complex, discovering a sparse code is useful to distill this big data into its most informative and representative parts for many applications like de-noising corrupted data, minimizing storage using compression, and tasks like segmentation and classification. Furthermore, sparse codes are an important ingredient in the application of accelerating signal acquisition. The main idea is that by transforming complex data into a sparse code, the number of measurements needed to reconstruct a full resolution signal becomes proportional to the number of elements in the sparse representation. Therefore, the sparser the code, the few measurements are needed, and the faster the acquisition.

Prior work have developed *angular* sparse coding for dMRI frameworks which

aim to sparsely represent diffusion signals in each voxel. However, the global sparsity is restricted by the number of voxels in the volume, therefore, limiting acceleration levels. We propose a joint *spatial-angular* sparse coding framework which overcomes these limitations to achieve levels of sparsity unattainable in prior formulations. We complete this chapter with a number of extensions to popular sparse coding algorithms which efficiently optimize over the large-scale size of dMRI data and are applicable to any similarly structured large scale data.

1.2.2 (k, q)-Compressed Sensing with Spatial-Angular Sparsity Priors

Next, in Chapter 4 we use the sparse codes discovered in Chapter 3 for the acceleration of dMRI using a signal processing paradigm known as *compressed sensing*, which permits the reconstruction of a signal using a sampling rate that is proportional to the sparsity of the representation with an adequate sensing scheme. One of the first applications of compressed sensing was the acceleration of MRI by subsampling in k-space, the frequency domain in which MRI images are acquired. Compressed sensing has also been applied to dMRI, where subsampling has taken place in q-space, the frequency domain analogue of the angular domain where diffusion signals are acquired. The state of the art has even combined prior works by subsampling jointly in the combined (k, q)-space. However, these frameworks rely on *angular* sparse coding.

In contrast, by discovering sparser codes from Chapters 3 within our *spatial-angular* framework, we aim to minimize the number of samples we need to reconstruct a full signal and accelerate dMRI acquisition to levels beyond the state of the art.

1.2.3 Spatial-Angular Dictionary Learning

Then, Chapter 5 is dedicated to optimizing the levels of sparsity by dictionary learning. In particular, a sparse code depends on what is known as a dictionary, consisting of a set of words or atoms that span the domain of the signal. The choice of this dictionary will dictate how sparse the representation of the data is. In Chapters 3 and 4 we will use existing dictionaries to produce a sparse spatial-angular representation of dMRI. Instead, in Chapter 5 we will optimize our choice of dictionaries by learning them directly from dMRI signals. This machine learning methodology, known as *dictionary learning*, uses examples of dMRI signals to train dictionaries that may naturally produce sparser codes than generic, or analytic, dictionaries. Like angular sparse coding, angular dictionary learning for dMRI has been well-studied, producing well-tailored dictionaries for real dMRI signals. The main contribution in this chapter is the extension to joint *spatial-angular* dictionary learning for dMRI. While existing dictionary learning algorithms can be applied to the structure of our problem, the major novelty and usefulness of our proposed framework to the machine learning community is a guarantee of global optimality which prior works have been unable to provide.

1.2.4 Future Directions: Convolutional Spatial-Angular Sparse Coding

One restriction of learning global spatial-angular dictionaries, however, is the need of many full size dMRI training example data sets which becomes computationally prohibitive. Therefore, as is common in image processing, instead of learning global spatial dictionaries, one can choose to learn smaller local dictionaries over many patches in an image. This greatly reduces the computational load of training but provides challenges for global reconstruction. A machine learning methodology to reconcile local patch-based spatial dictionaries with global reconstruction is known as *convolutional* sparse coding and dictionary learning. In Chapter 6, we will present a preliminary framework for *convolutional spatial-angular* sparse coding for patch-based dictionaries.

Chapter 2

Background

In this chapter we provide background material and mathematical notations necessary for better understanding the contributions of this thesis and their place within the larger context of the field. We first give a brief summary of the principles of diffusion MRI in Section 2.1, including an overview of clinical applications, the physics of the imaging modality, models of diffusion, and methods for acceleration of the acquisition.

Then, in Section 2.2, we detail the mathematical notions of sparse reconstruction with background and literature review on the general problems of sparse coding, compressed sensing, and dictionary learning. Applying these topics to diffusion MRI and adapting them to the unique structure of our data will be the aim of this thesis.

2.1 Principles of Diffusion MRI

2.1.1 Overview and Clinical Applications

Diffusion magnetic resonance imaging (dMRI) is a medical imaging modality used to reconstruct the anatomical network of fiber bundles in the brain, *in vivo*. Since its inception in 1985 [16] with the world's first diffusion magnetic resonance image, there has been a huge increase in the number of major scientific discoveries related to brain anatomy and the investigation of neurological disorders and diseases. As a notable example, dMRI is being used to identity biomarkers to predict the early onset of Alzheimer's and Parkinson's diseases related to the structural degradation of the neuronal connections that may cause memory and motor deficiencies [2, 17, 18].

Other applications of dMRI include stroke [19], dementia [20] schizophrenia [21], and autism [22]. dMRI is also being used at the forefront of traumatic brain injury research [23] related to military veterans and most recently in the National Football League in the United States [24]. As water diffusion can also be measured in the body, whole body dMRI has been used for applications including cancer research [25], as well as for modeling of heart contractions and arrhythmia [26] and even tongue lesions [27].

As diverse imaging technologies become increasing available for single subjects, researchers are able to combine dMRI with other modalities like functional MRI (fMRI) [28] and electroencephalography/magnetoencephalography (EEG/MEG) data [29] for

multi-modal studies relating anatomy to brain function. Another important multi-modal application of dMRI is neuro-surgical planning [30]. All of these applications, require a series of algorithms for processing dMRI data. Figure 2.1 summarizes the typical pipeline for dMRI processing. The following is a overview of the main steps in the pipeline as well as many open research challenges.

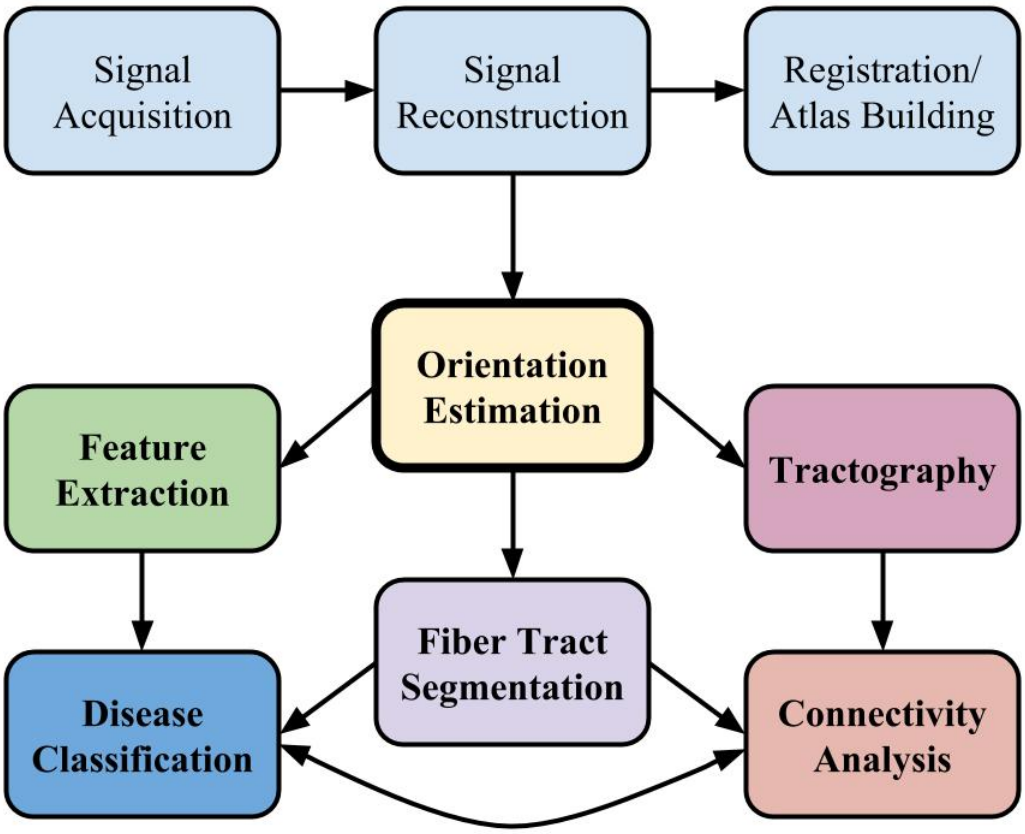


Figure 2.1: Overview of the dMRI pipeline. Fiber orientation estimation is the focal point of dMRI for which all other down-stream processes and analyses depend on. Because fiber orientation estimation occurs at the voxel-level, a voxel-centered viewpoint permeates to all other processes and applications in dMRI. (While this is a simplistic schematic, there are numerous other connections between tasks not shown here for which algorithms benefit from their mutual information, such as registration combined with orientation estimation, or fiber tract segmentation combined with tractography.)

Diffusion Orientation Estimation. Unlike other imaging modalities like MRI, fMRI, CT, and PET, dMRI has the unique ability to image the neuroanatomy by measuring water diffusion in the brain. Since water diffuses preferentially along the path of structural boundaries, like neuronal fiber bundles, by measuring the degree of local water diffusion, researchers can infer the spatial orientations of fiber bundles in each voxel of the brain [31]. These orientation estimations are mathematically represented as 3D probability distribution functions (PDFs), the peaks of which indicate the most probable directions of fibers in a voxel.

Figure 2.2 shows a field of these PDFs, sometimes known as orientation distribution functions (ODFs), with fiber tract streamlines propagated by following their peak directions. For a real brain example, Figure 2.3 shows the ODFs in each voxel of an slice of a human brain dMRI. As is evident from these images, fiber populations can exhibit very intricate configurations such as crossing, bending, twisting and fanning at voxel-level resolutions. Therefore, developing accurate models of diffusion to estimate the PDFs is an important first step for subsequent processing and applications. Diffusion estimations can be refined by de-noising and smoothing methods that also incorporate diffusion information from a spatial neighborhoods [32] for example.

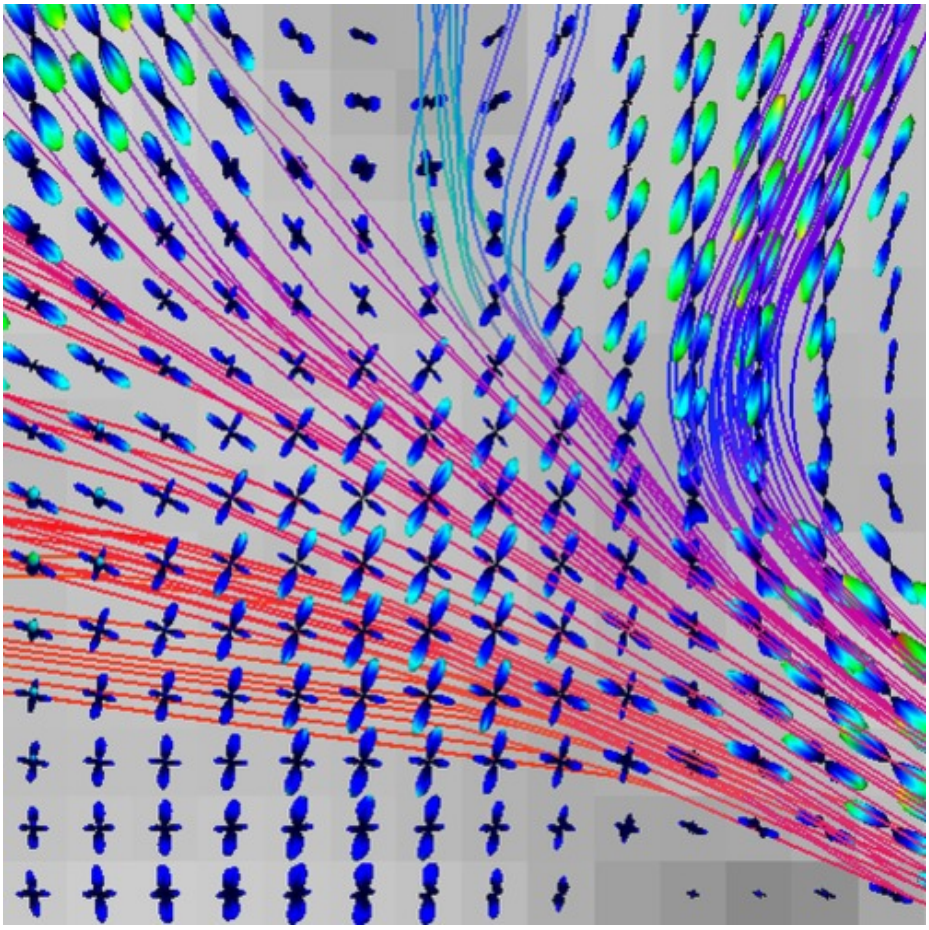


Figure 2.2: Field of orientation distribution functions (ODFs) indicating the most probable direction of a fiber populations. Following their peak directions, fiber tract streamlines are traced to reconstruct the neuro-anatomy.[Image adapted from Institute for Numerical and Applied Mathematics, Universary Gttingen.]

Feature Analysis. One immediate application of these diffusion models is to derive and extract small sets of scalar features that can reveal important information about complex dMRI data such as PDF shape descriptors or quantitative characteristics of diffusion that can indicate the integrity of fiber connections. The task of reducing these complex models to a few scalar quantities, called *feature extraction*, is beneficial for clinicians to view the high-dimensional data in an interpretable manner.

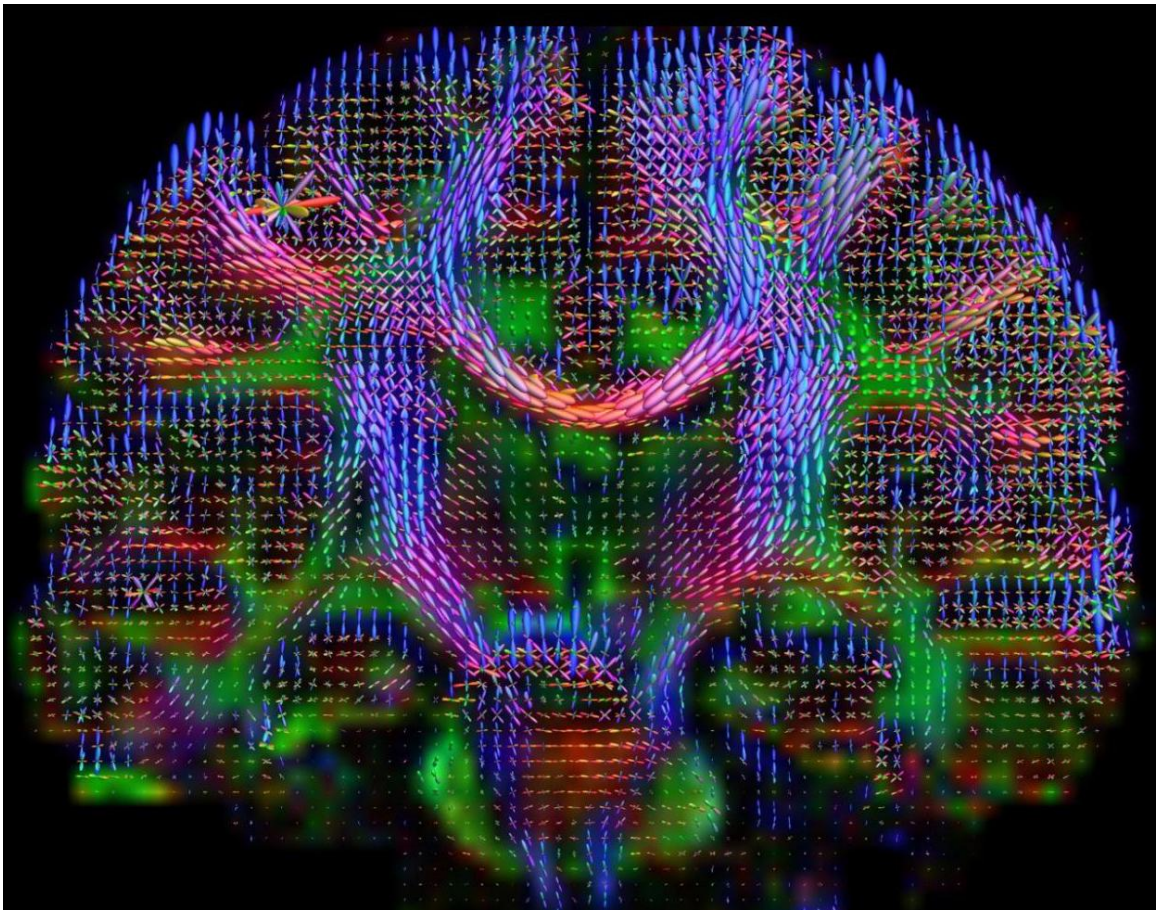


Figure 2.3: Orientation distribution functions (ODFs) estimated for each voxel of a brain volume indicate the direction of fiber tracts. Colors reveal directional information for visualization. There exist strong single-fiber ODFs in strong tracts like the corpus collosum connecting the left and right hemispheres (center red U shaped tract), and two-fiber and three-fiber crossing ODFs at the intersection of fiber tracts. [Image adapted from Walt Schneider/University of Pittsburgh]

For large-scale clinical studies, these features can then be used to statistically evaluate and compare populations of subjects in order to identity biomarkers that may be indicative of certain disease attributes. In addition, the reduction of high-dimensional data to low-dimensional representations is useful for machine learning tasks like disease classification which allows researchers to train models on feature

data and build classifiers that can distinguish between normal and abnormal subjects and predict if a new subject may be a candidate for disease. Figure 2.4 gives an example of a scalar feature called fractional anisotropy, which measures the integrity of fiber connectivity.

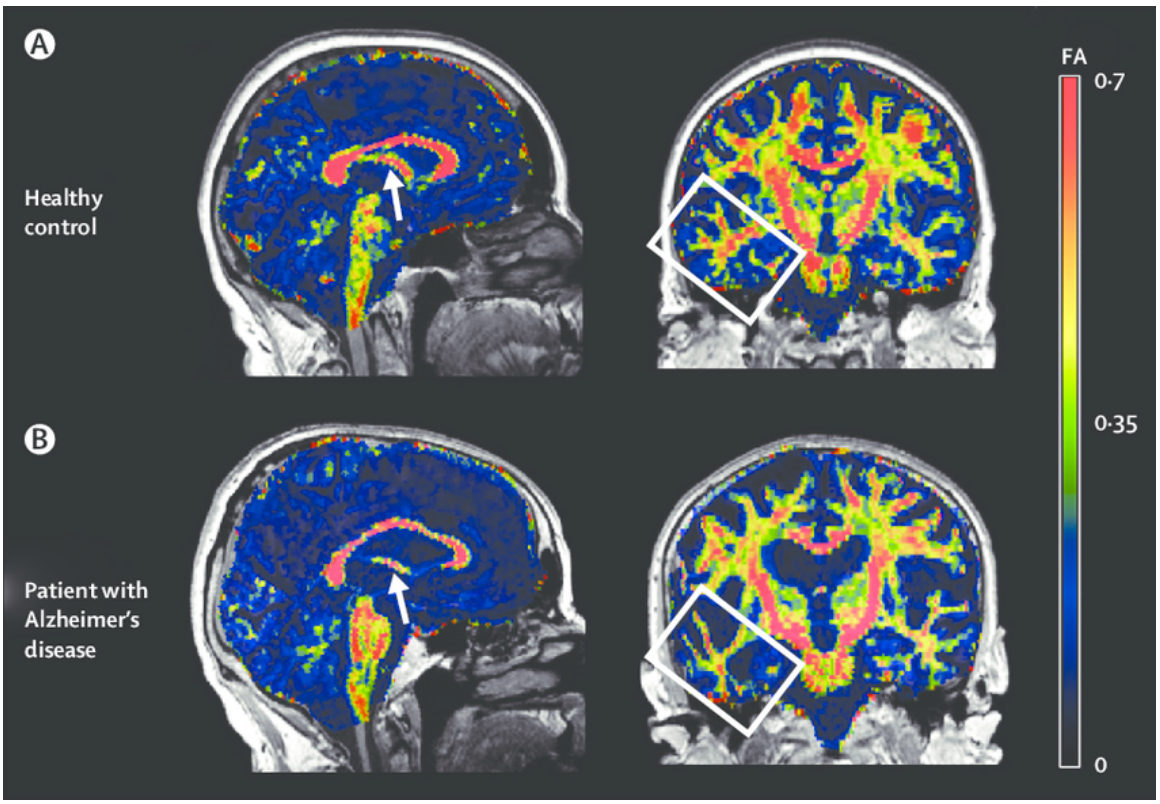


Figure 2.4: Analysis of fractional anisotropy (FA) between healthy control and a patient with Alzheimer's disease. FA indicates the integrity of neural connectivity, and a decrease in FA is evident in the limbic system (white arrow, left), the memory center of the brain, and the temporal lobe (white box, right), known to deteriorate in patients with dementia. [Image adapted from [2].]

Registration and Atlas Building. The shape of every brain is different and subjects scanned at different sites will have diverse imaging parameters and conditions. In order to normalize medical imaging data, a task known as *registration* is used to

transform individual subjects to a common space, known as an *atlas space*. An atlas is an anatomical map or blueprint, with sometimes labeled parts, that is constructed from a large population of healthy subjects (see Figure 2.5). An atlas serves as an estimation of a population "average" with respect to individual subjects. New subjects are then compared to the atlas in a statistically valid manner to understand if their pathology is close to average or abnormal and how their anatomy differs. To do this, the task of registration aims to find a transformation between two images that minimizes their differences.

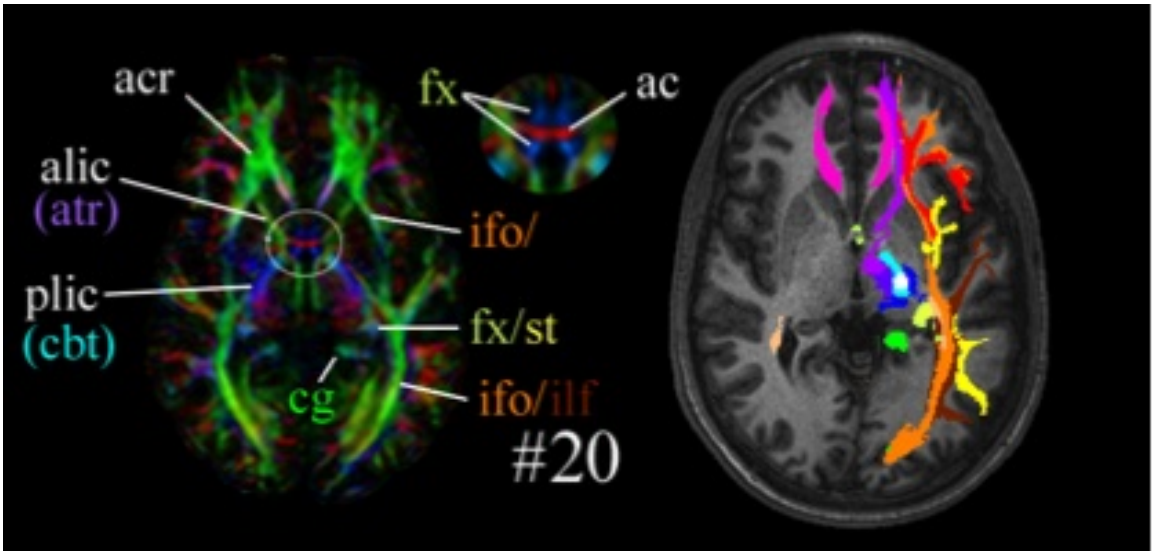


Figure 2.5: Example of a dMRI atlas with labeled tracts and anatomical regions of interest. Labeled atlases can be constructed from populations of imaging data as an average or standard brain anatomy. An atlas can be registered to a new subject to label and segment regions of interest. [Image adapted from [3].]

For instance, if one subject accidentally rotated their head 15 degrees to the left within an MRI scanner, one can rotate the image data 15 degrees back to the right to compare with another normally orientated image in the same anatomical space.

This registration consists of a simple rigid rotation transformation. In practice, more complicated transformations are needed to align different brain shapes requiring a smooth nonlinear deformation of the image to align with an atlas. While much work has been devoted to the registration of natural and medical images, the registration of dMRI, is more involved due to the complex geometric structure of diffusion signals. In particular, a key challenge is that a transformation of the spatial domain produces a reorientation of the angular domain of the diffusion signals, making it more difficult to define a suitable registration objective.

Fiber Tract Segmentation. Another use of atlases is for the task of *segmentation*, or the identification and labeling of objects or structures in an image into separate parts or categories. Segmentation of anatomical structures is an important problem in medical image analysis for various applications like surgical planning, lesion detection, and volumetric analyses like estimating changes in cortical thickness of the brain, monitoring temporal images of the heart, or comparing the shapes of organs in a population. In the past, images were segmented manually, like the white matter, grey matter, and cerebrospinal fluid regions in the brain. An ongoing research challenge is developing algorithms for automatic segmentation to eliminate the need for thousands of man-hours needed for manual segmentation. One approach is to manually label an atlas (e.g., the volume for one subject), register the atlas to the volume of an unlabeled subject, and then transfer the labels from the atlas to the unlabeled volume using the obtained registration map.

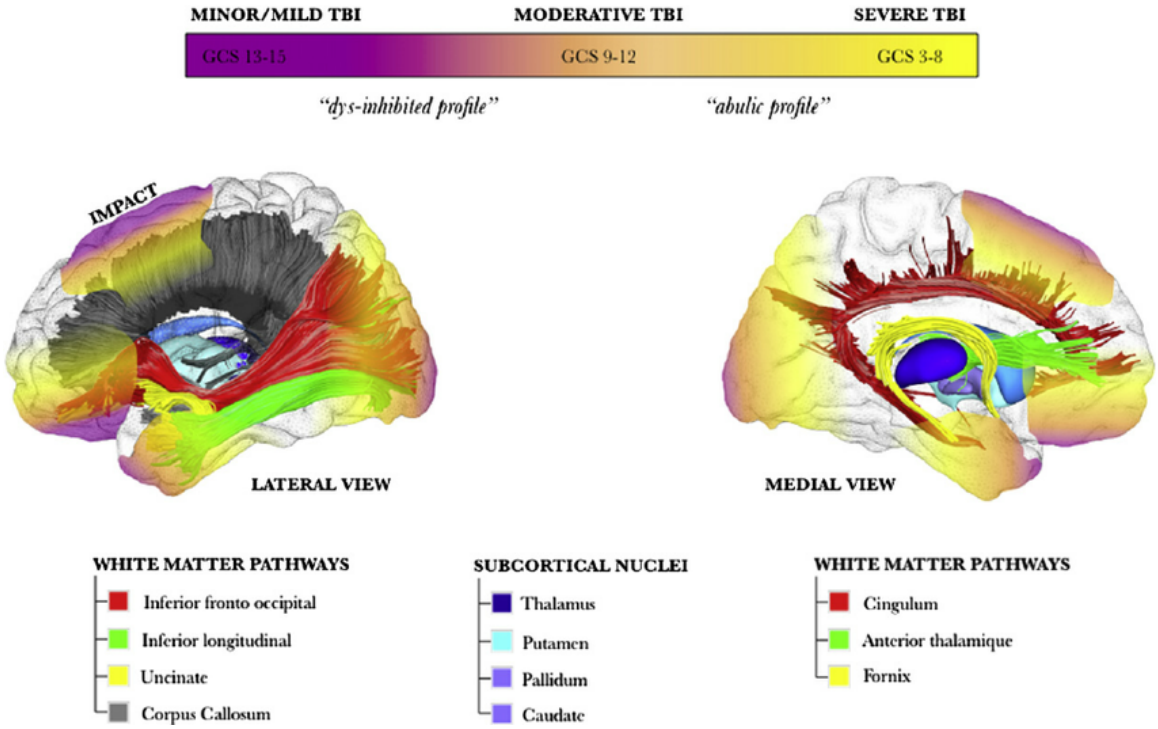


Figure 2.6: Fiber tract segmentations used in the analysis of traumatic brain injury (TBI). [Image adapted from [4].]

In the case of dMRI, researchers are interested in segmenting different groups of fiber tracts which can reveal connectivity and shape information about neurological diseases and disorders, hence typical labels in the atlas correspond to different white matter pathways, as illustrated in Figure 2.6. Another approach is to use unsupervised learning techniques, such as *clustering*, to group voxels that have similar diffusion information as belonging to the same fiber tract [33]. Then tract shapes and volumes can be analyzed to better understand the integrity of different anatomical connections. This is an example of a voxel-wise vantage point that incorporates spatial information to achieve a task.

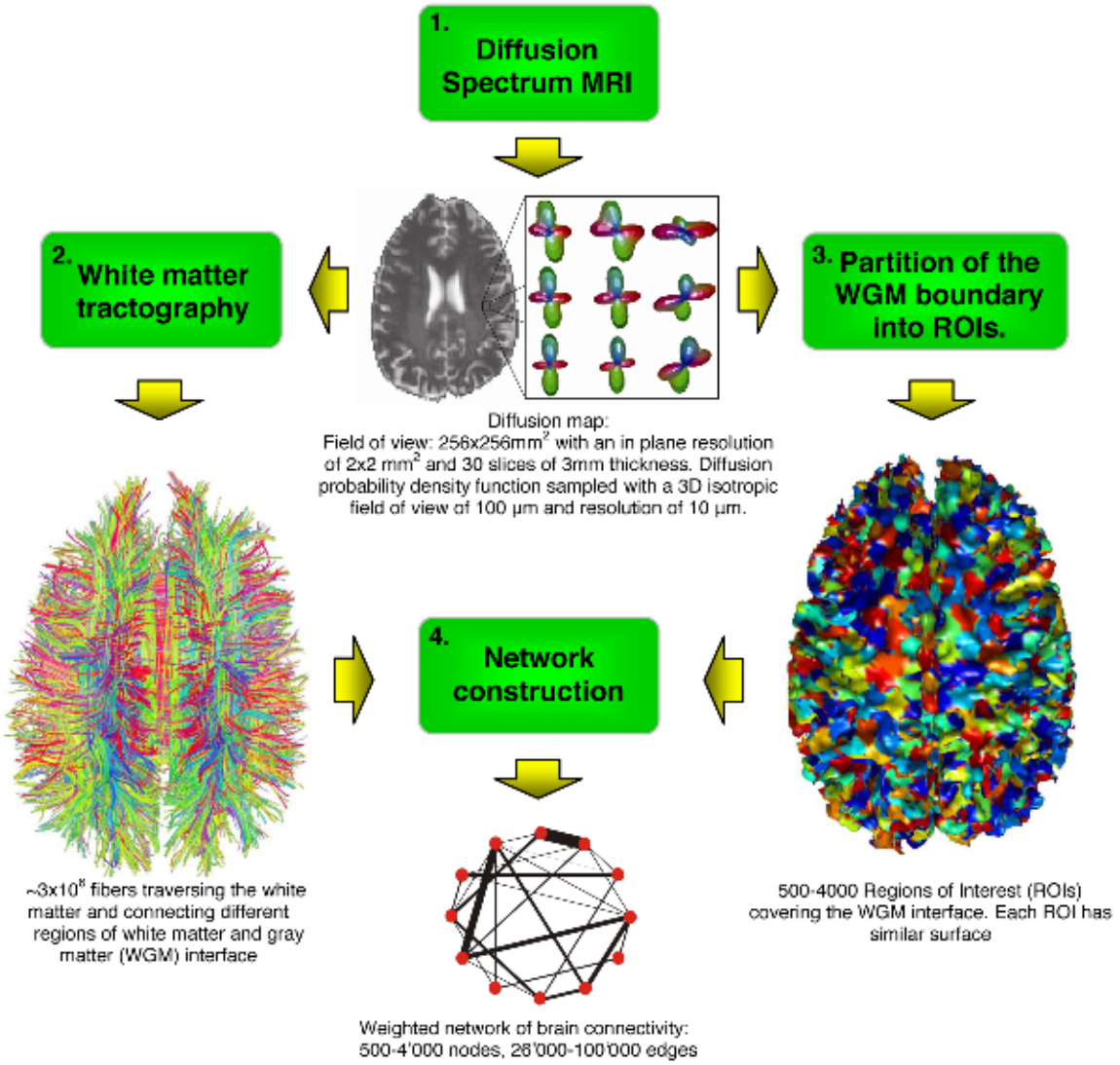


Figure 2.7: Overview of connectivity network construction based on fiber tractography and anatomical brain region partitioning. From orientation distribution function estimation, fiber tracts are reconstructed. Then from a dMRI atlas, labeled brain regions are segmented. A graphical network of connectivity between brain regions is constructed by identifying the tracts that link various brain regions. Statistical population analyses can be run on this graphical model. [Image adapted from [5].]

Tractography. One of the end-goals of dMRI is known as fiber tractography, which is concerned with the reconstruction of a network of fibers connecting different regions in the brain for example. One possible approach is to follow the peak directions of

PDFs within each voxel. This first involves finding the peaks of a distribution over a continuous domain, which can be challenging when the distribution is multimodal, as is for example in the case of crossing fibers. Then, with these orientations detected for each voxel, tractography starts at a set of seed locations and looks to neighboring voxels which are contained in cones emanating from the peak directions of the seed voxel. This process repeats for the neighboring voxel and a streamline is propagated, connecting voxels in various parts of the brain volume.

There have been two classes of methodologies for tractography, deterministic and probabilistic [34] based on a degree of uncertainty of the peak directions. As the fiber tracts are populated over an entire volume, one can extract features like the number of streamlines that connect two regions of interest or the end to end lengths of fiber tracts, each of which give an indication of the integrity or strength of a connection. These fiber tracts can then be integrated with labeled brain regions or fMRI data revealing how different functional parts of the brain interact anatomically. Furthermore, we can compare the fiber connections of healthy and diseased subjects to understand differences in wiring and strength in connectivity.

Connectomics. From a different viewpoint, the connections obtained by tractography can be analyzed through the lens of the mathematical field of graph theory, where the nodes of the graph are local brain regions and the edges are the fiber tracts connecting each brain region. With an atlas of segmented brain regions, graphs can be compared over a population of dMRI subjects (see Figure 2.7). Running statistical

models over these graphs is an open research challenge, especially where the sizes of the graphs run into the terabytes[1]. The complete set of connections are known as the human connectome[2] and updating our knowledge of the wiring of the human brain is an active and exciting area of research.

To better understand how to represent the structure of dMRI data, it is important to understand how the signals are acquired. In the next section, we provide a background on the physics of dMRI, starting with first principles of magnetic resonance imaging in Section 2.1.2.1 and then building to diffusion imaging in Section 2.1.2.2. We will then discuss the many models for sampling diffusion data and the sets of features which characterize diffusion properties in Section 2.1.3. Then in Section 2.1.4, we will discuss the process of reconstructing dMRI signals and estimating orientation distribution functions. We will end with an important discussion of recent methods that aim to accelerate the dMRI acquisition with the goal of increasing the clinical applicability of this emerging technology.

2.1.2 Physics of dMRI

2.1.2.1 MRI Acquisition (k-space)

Magnetic resonance imaging (MRI) gets its name from the excitation of magnetic dipoles, or spins, of hydrogen nuclei in the body. Since the body, and in particular

[1]For cutting-edge methods and open-source code, visit https://neurodata.io/
[2]Visit the Human Connectome Project at http://www.humanconnectomeproject.org/

the brain, is mostly made up of water molecules, MRI systems aim to measure and differentiate the concentrations of water in various brain tissues by exciting magnetic dipoles with powerful magnetic fields. This allows the system to distinguish between white matter, grey matter, and cerebral fluid in the brain, for example, as well as lesions or other foreign structures in the body.

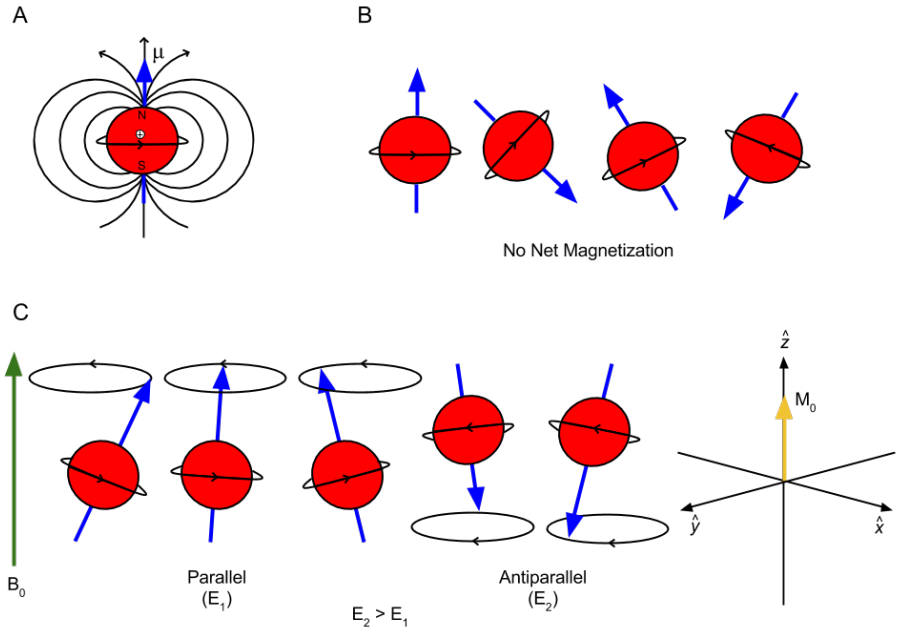


Figure 2.8: Physical properties of the spin of protons in presence of a magnetic field. A. Protons are charged objects with magnetic dipole moment, or the magnitude of the magnetic field, given by μ. B. In the absence of an external magnetic field, each proton is free to spin in any orientation. C. With a large magnetic field $\vec{\mathbf{B}}_0$, the protons orient (either in parallel or antiparallel) to the direction of the field. Because the energy state of the parallel orientation (E_1) is less than that of antiparallel (E_2), more protons will orient in parallel cause a net magnetization $\vec{\mathbf{M}}$ in the direction of $\vec{\mathbf{B}}_0$. [Image adapted from [6].]

In the absence of any external magnetic field, dipoles are free to spin along their individual axes pointing in random directions. When an external magnetic field is introduced, which we denote by the vector $\vec{\mathbf{B}}_0$, the dipoles exposed to this field will

align with the field's direction and spin around this axis at a frequency known as the Larmor frequency (see Figure 2.8). The dipoles orient to the axis of the magnetic field by choosing one of two paths, spin-up (parallel) or spin-down (anti-parallel). Since the easier of the two options, or path with lower energy, is parallel to $\vec{\mathbf{B}}_0$, more dipoles will align this way than anti-parallel, generating a net magnetic field $\vec{\mathbf{M}}$ in the spin-up direction. The dipoles cannot align exactly, however, and so they precess around the axis like a gyroscope. Therefore $\vec{\mathbf{M}}$ has non-zero longitudinal ($\| \vec{\mathbf{B}}_0$) and transverse ($\perp \vec{\mathbf{B}}_0$) vector components, $\vec{\mathbf{M}}_L$ and $\vec{\mathbf{M}}_T$, respectively.

Once in an aligned precession, a radio frequency (RF) pulse equal to the Larmor frequency is sent to excite the dipoles, causing them to tilt their axes into the plane perpendicular to $\vec{\mathbf{B}}_0$ and precess in the plane (known as a 90° pulse). The dipoles will rotate at frequency ν, given by the Larmor equation $\nu = \gamma B_0$, where γ is the gyromagnetic ratio (a property of water protons) and B_0 is the magnitude of the magnetic field vector $\vec{\mathbf{B}}_0 = B_0\vec{\mathbf{u}}$. When the RF pulse is removed, the dipoles will relax to their original, stable energy states in alignment again with $\vec{\mathbf{B}}_0$, giving off excess energy which is then detected by receiver coils in the MRI system.

As the population of dipoles are excited and relaxed, the net magnetization $\vec{\mathbf{M}}$ orients perpendicular to $\vec{\mathbf{B}}_0$ and reorients back. The time it takes to relax to its original state is indicative of the type of tissue the water molecules live in. As such, there are two main quantities of $\vec{\mathbf{M}}$ that the MRI coils measure in the relaxation stage:

1. T_1: The time (ms) it takes $\vec{\mathbf{M}}_L$ to relax to a state of 63% of its original magnitude.

2. T_2: The time (ms) it takes $\vec{\mathbf{M}}_T$ to relax to a state of 63% of its magnitude at the current excited state.

Each biological tissue has different T_1 and T_2 values and the MR images that are formed by these quantities are known as T_1-weighted and T_2-weighted MR images. An additional 180° RF pulse is applied to bring the protons back into phase after the initial relaxation.

But first, the process of going from measurements to images requires spatial localization. This localization is achieved by small perturbations to the magnetic field, or gradient pulses for short amounts of time, in the x, y, and z directions, denoted $\vec{\mathbf{G}}_x, \vec{\mathbf{G}}_y$, and $\vec{\mathbf{G}}_z$. First, $\vec{\mathbf{G}}_z$ (pointing parallel to $\vec{\mathbf{B}}_0$) is applied which selects a slice orthogonal to $\vec{\mathbf{B}}_0$. After slice selection, $\vec{\mathbf{G}}_y$ and then $\vec{\mathbf{G}}_x$ are applied to encode the phase and frequency, respectively, resulting in unique slice, phase and frequency coordinates. The ordered combination of these pulses with the original RF pulse results in what is known as a pulse sequence to form an MR image (see Figure 2.9).

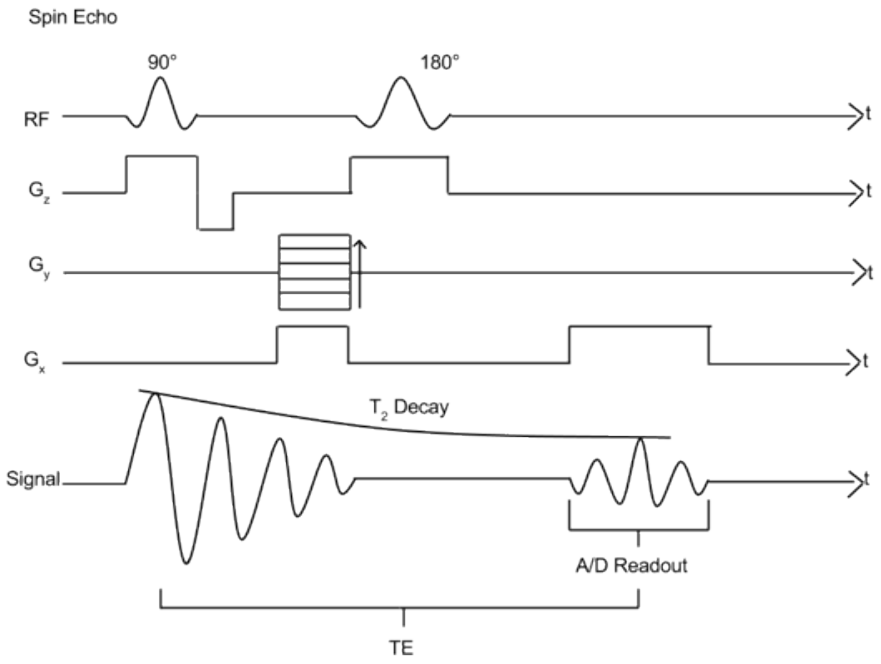


Figure 2.9: MR pulse sequence. The RF undergoes two flips at 90° and 180°. $\vec{\mathbf{G}}_z$, $\vec{\mathbf{G}}_y$, and $\vec{\mathbf{G}}_x$ are the slice selection, phase encoding, and frequency coding gradients, respectively. T_2 is the time it takes for the magnitude of the signal to relax to 63% of its original magnitude. The readout of the signal occurs during the spin echo after time TE.

For each slice, measurements taken by an MRI scanner are in the complex domain of k-space (k_x, k_y) given by the following relationship with the localized gradients $(\vec{\mathbf{G}}_x, \vec{\mathbf{G}}_y)$ as a function of time t:

$$k_x(t) = \frac{\gamma}{2\pi} \int_0^t G_x(\tau) d\tau \qquad \text{and} \qquad k_y(t) = \frac{\gamma}{2\pi} \int_0^t G_y(\tau) d\tau \tag{2.1}$$

Finally, from the k-space signal, an MR image in x-space is formed by taking the inverse 2D Fourier Transform (see Figure 2.10).

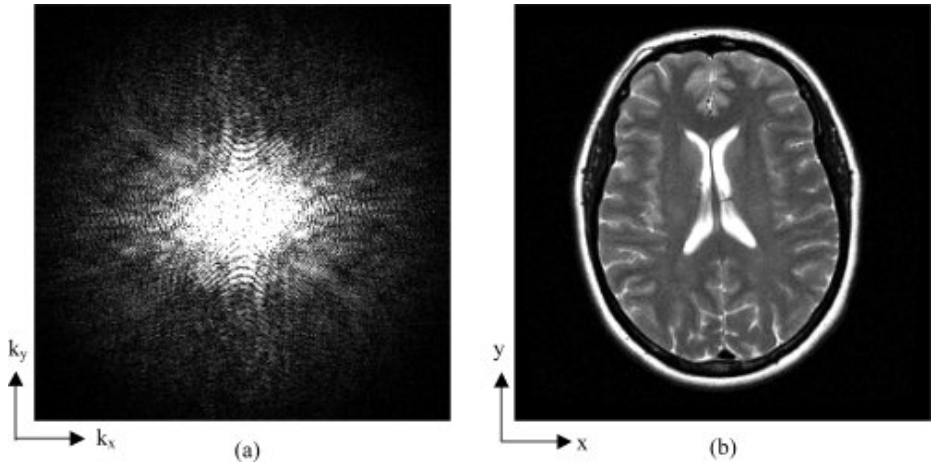


Figure 2.10: MRI measurements in k-space frequency domain (a) related to x-space image domain (b) by the 2D Fourier Transform. [image adapted from [7]

2.1.2.2 dMRI Acquisition ((k, q)-space)

Water Diffusion. The basis of diffusion MRI is the measurement of water diffusion, or flow, in the brain. Water diffusion can be characterized as the displacement of a water molecule over a certain amount of time. The farther the water travels in a given time frame, the greater the degree of diffusion in that direction.

Formally speaking, let $\mathbf{R}_0 \in \mathbb{R}^3$, be the initial position of a water molecule, and $\mathbf{R}_\tau \in \mathbb{R}^3$ be the final position after time τ. The displacement vector $\mathbf{r} \in \mathbb{R}^3$ is given by $\mathbf{r} = \mathbf{R}_\tau - \mathbf{R}_0$. In fact, we can only measure the positions at certain times and not the path of a water molecule itself since water molecules will follow a random trajectory, or random walk, according the laws of Brownian motion (see Figure 2.11). Therefore, it is more informative to measure the diffusion process of an ensemble, or population, of water molecules, and track the average displacement, given by the

quantity $\langle \mathbf{r}\mathbf{r}^\top \rangle$. Einstein [35] discovered that, in the absence of structural boundaries when water is free to move in all directions without constraint, $\langle \mathbf{r}\mathbf{r}^\top \rangle$ is proportional to the displacement time τ by the relation

$$\langle \mathbf{r}\mathbf{r}^\top \rangle = 6D\tau \tag{2.2}$$

where D is known as the diffusion coefficient.

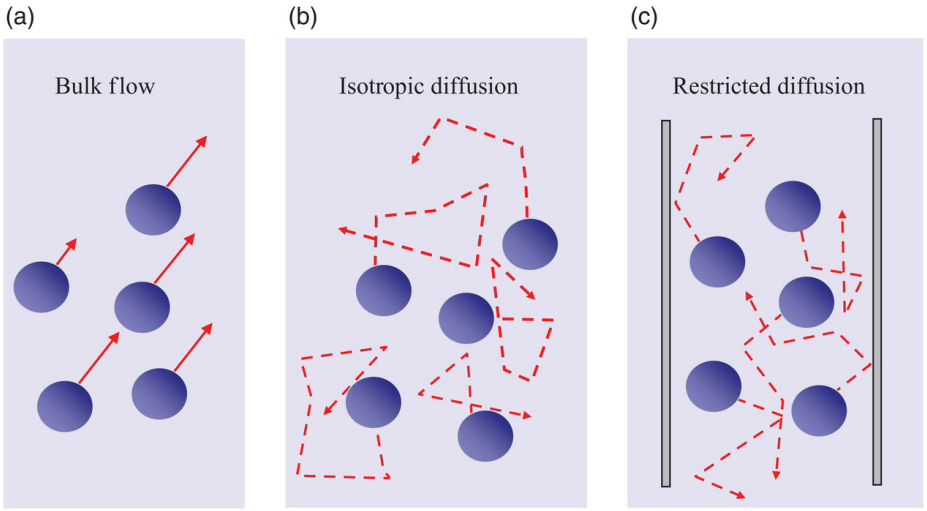


Figure 2.11: Water diffusion by Brownian motion. (a) Normal directional flow of water. (b) Water molecules are free to travel in all directions following Brownian motion (isotropic). (c) In the presence of a restrictive boundary, water will travel with higher probability along that orientation (an-isotropic). [Image adapted from [8].]

In this setting of an unconstrained environment, diffusion is called isotropic (coming from the Greek roots *iso* meaning 'equal' and *tropos* meaning 'ways'). Alternatively, in the presence of obstructions constraining the water displacement, known as anisotropic diffusion, Einstein's equation does not apply and the diffusion coefficient must be estimated in another way. By measuring the average water displacement

in a set of directions, diffusion imaging systems aim to estimate the coefficients of diffusion, known as the apparent diffusion coefficient (ADC), in all directions in order to paint a picture of the underlying structures present in various parts of the brain. Building on the pulse sequence of MRI systems, additional magnetic fields are applied to capture this diffusion information, which we discuss in detail in the next section.

Diffusion Weighted Imaging (DWI). Diffusion weighted images (DWIs) are a variation of MR images with an application of additional magnetic fields applied along a direction, known as a diffusion gradient direction, given by

$$\vec{\mathbf{g}} = \vec{\mathbf{B}}_0 + \vec{\mathbf{G}}_x + \vec{\mathbf{G}}_y + \vec{\mathbf{G}}_z. \tag{2.3}$$

Each MR image acquired with a different gradient direction, is known as a DWI (see Figure 2.12). By taking a series of DWIs, we can begin to investigate the diffusion landscape in 3D space. As we will show below, these gradient direction measurements live in what's called q-space, an analogue of k-space in the angular diffusion domain. Since each DWI is acquired in k-space and weighted by the measurement in q-space, the total combined space of measurements for dMRI is in the product (k, q)-space in $\mathbb{R}^3 \times \mathbb{R}^3$.

The vector $\vec{\mathbf{g}}$ is actually bipolar, consisting of two components $\vec{\mathbf{g}}_+$ and $\vec{\mathbf{g}}_-$ with the same direction and magnitude but opposite orientations. After the initial RF pulse, the first gradient $\vec{\mathbf{g}}_+$ is applied which adds a positive phase to each spin. Then

$\vec{\mathbf{g}}_-$ is applied adding the opposite phase to each spin. The successive application of these two bipolar gradients is able to detect changes in the average spins of the water molecules. If the molecules return to their original average spin direction after both gradients, this implies no displacement has taken place. On the other hand if the final spins are not in phase, then they have been displaced according to the random walk Brownian motion of water diffusion. This difference in field strength is then detected by a decrease in the relaxation of $\vec{\mathbf{M}}_T$ (corresponding to T_2-weighted imaging).

Stejskal and Tanner [36] invented the pulse sequence known as the pulse gradient spin echo (PGSE) sequence, which applies these short duration diffusion gradient pulses $\vec{\mathbf{g}}_+$ and $\vec{\mathbf{g}}_-$, as described. Then, defining $\mathbf{q} = \gamma\delta\vec{\mathbf{g}}$, where δ is the short duration of each diffusion gradient pulse, the diffusion signal at gradient direction $\mathbf{q}$ is given by the equation

$$S(\mathbf{q}) = S_0 e^{(-b(\mathbf{q})D)} \tag{2.4}$$

where S_0 is a function of T_1, T_2, and other sequence parameters, D is the ADC, and

$$b(\mathbf{q}) = (\Delta - \frac{\delta}{3})||\mathbf{q}||^2 \tag{2.5}$$

known as the b-value, with Δ the time between pulses $\vec{\mathbf{g}}_+$ and $\vec{\mathbf{g}}_-$. The b-value is an important parameter, based on the length of the $\mathbf{q}$ vector, that effects the signal-to-noise ratio (SNR). Larger b-values result in noisier data measurements and a poor SNR, while smaller b-values provide lower signal attenuation. In particular, when

$b = 0$, $S(\mathbf{q}) = S_0$, and so the baseline S_0 is commonly known as the $b = 0$ or b_0 image. The units of the b-value are s/mm^2 and may range from about 500-5000 s/mm^2.

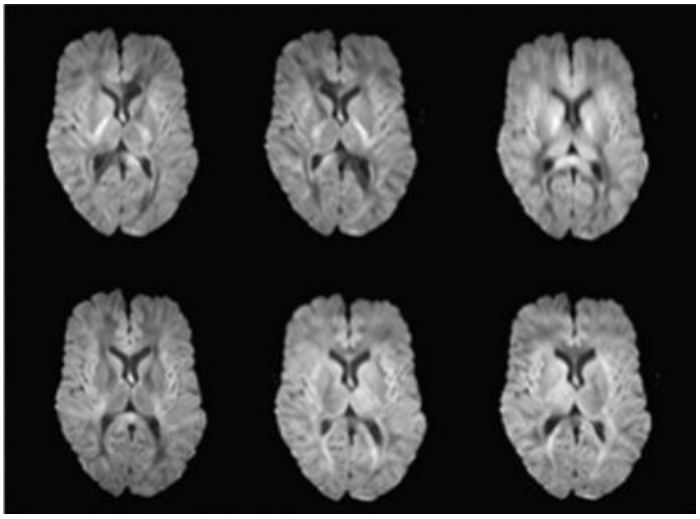


Figure 2.12: DWIs each with a different diffusion weighting measured in q-space. [Image adapted from [9].]

In (2.4) one could simply solve for D for each value of $\mathbf{q}$ and get a scalar quantity of diffusion for each gradient direction, which has been shown as a useful feature of the diffusion process in many early works [37]. However, this value may not be enough to characterize the anatomical intricacies of the diffusion process and therefore, researchers have aimed to better understand a fuller explanation of diffusion as a 3D probability distribution function over all directions in 3D space.

More concretely, consider the quantity $P(\mathbf{R}_\tau|\mathbf{R}_0, \tau)$, the probability of a water molecule at position $\mathbf{R}_0$ being displaced to $\mathbf{R}_\tau$ over time τ. As stated previously, the displacement of one water molecule within Brownian motion is difficult to measure accurately, hence we consider an ensemble of water molecules in a given voxel do-

main Ω_v and estimate their average propagation over time τ, known as the Ensemble Average Propagator (EAP):

$$P(\mathbf{r}|\tau) = \int_{\mathbf{R}_0 \in \Omega_v} P(\mathbf{R}_\tau|\mathbf{R}_0, \tau)\rho_v(\mathbf{R}_0)d\mathbf{R}_0 \tag{2.6}$$

where $\rho_v(\mathbf{R}_0)$ is a density of molecules in voxel v with $\int_{\mathbf{R}_0 \in \Omega_v} \rho_v(\mathbf{R}_0)d\mathbf{R}_0 = 1$.

Many works have recently been studying the effects of time τ [38, 39], but in the most classical setting, we assume τ to be constant for every voxel and so the EAP can reduce to $P(\mathbf{R})$. Writing $E(\mathbf{q}) := \frac{S(\mathbf{q})}{S_0}$, when $\delta << \Delta$, the normalized signal is known to have a Fourier relationship with the EAP:

$$E(\mathbf{q}) = \int_{\mathbf{r} \in \mathbb{R}^3} P(\mathbf{r})\exp(-2\pi i \mathbf{q} \cdot \mathbf{r})d\mathbf{R} \tag{2.7}$$

Therefore the EAP can be computed by taking the inverse Fourier Transform of the normalized dMRI signal

$$P(\mathbf{r}) = \int_{\mathbf{q} \in \mathbb{R}^3} E(\mathbf{q})\exp(2\pi i \mathbf{q} \cdot \mathbf{r})d\mathbf{q}. \tag{2.8}$$

In practical MRI systems, calculating a continuous $P(\mathbf{r})$ for all $\mathbf{q} \in \mathbb{R}^3$ is infeasible and so accurate estimations of $P(\mathbf{r})$ depend on how one samples q-space. In the following sections, we outline various popularly used models for estimating $P(\mathbf{r})$ and other informative features of the diffusion process which can reveal important

structural properties of the brain anatomy for clinical applications of disease analysis and biomarker discovery.

2.1.3 dMRI Diffusion Models

Much effort has been spent deriving accurate models for reconstructing diffusion signals given noisy measurements and estimating the probability of diffusion $P(\mathbf{r})$ in a given voxel. By developing these diffusion models at each voxel, researchers can reconstruct anatomical fiber tracts of the brain *in vivo*. Figure 2.13 illustrates fibers crossing within a voxel and the orientations that we aim to estimate with models of probability. The following sections are devoted to deriving these diffusion models which each require their own imaging protocols for sampling q-space.

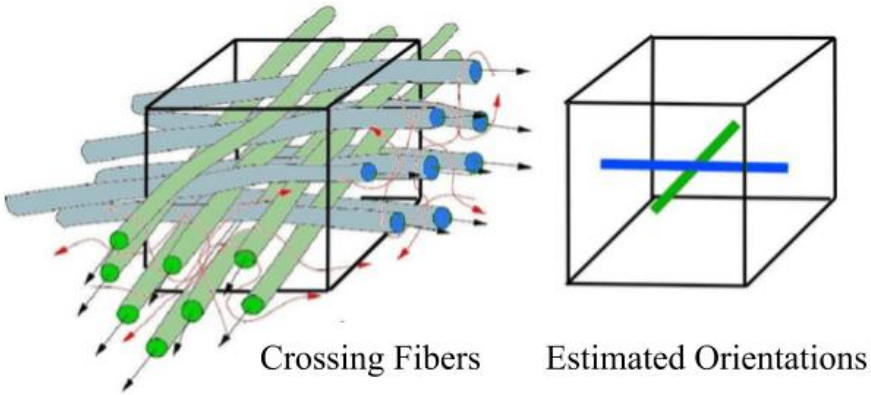


Figure 2.13: Fiber populations crossing within the space of a single voxel. The goal is to estimate the orientations of each fiber from measurements of water diffusion indicated by the green, blue and red arrows.

2.1.3.1 Diffusion Spectrum Imaging (DSI)

The most general and full diffusion model is known as diffusion spectrum imaging (DSI) [40] which estimates the EAP directly using the 3D Fourier relationship in (2.8) by densely sampling q-space on a 3D Cartesian grid (see Figure 2.14) with an range of b-values. The most common scheme is an $11 \times 11 \times 11$ grid that is contained within a ball with a radius of 5-lattice points, resulting in a total of 514 q-space samples. Adding the required $b = 0$ measurement gives 515. The EAP itself is evaluated on a discrete grid, giving a probability at a dense set of points $\{r_i\}_{i=1}^{M} \in \mathbb{R}^3$. However, the sheer number of DWIs required by DSI to estimate the EAP, prohibits it from being used in clinical practice. To reduce the number of q-space measurements needed, researchers have devised more simplified models which do not intend to estimate the full EAP, but a probability of diffusion restricted to the unit sphere. Such models are described next.

2.1.3.2 Diffusion Tensor Imaging (DTI)

Diffusion tensor imaging (DTI) [41] is one of the most well known dMRI techniques due to its simplicity. The major simplification of DTI is that it assumes a 3D Gaussian model of diffusion which allows for the computation of the Fourier transform to be trivial. More specifically, in DTI,

$$p(\mathbf{r}) = \frac{1}{\sqrt{(4\pi\tau)^3|\mathbf{D}|}} \exp\left(\frac{1}{4\tau}\mathbf{r}^\top \mathbf{D}^{-1}\mathbf{r}\right) \tag{2.9}$$

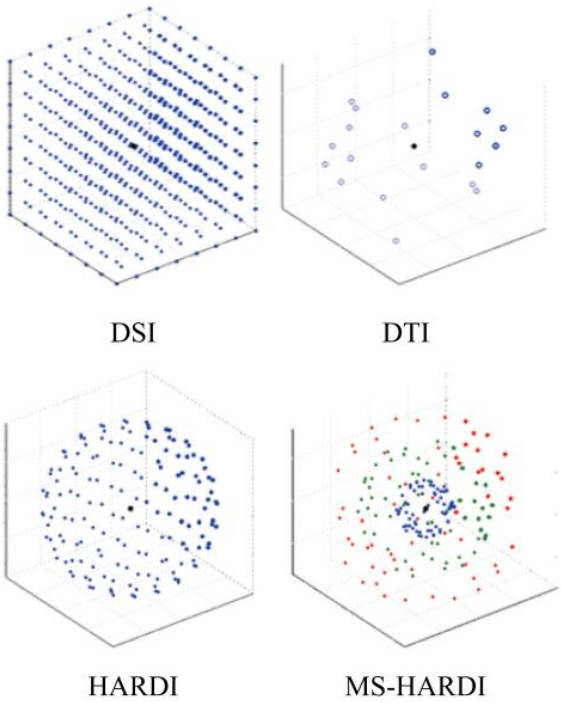


Figure 2.14: Sampling schemes in q-space for DSI (dense Cartesian grid), DTI (sparse unit sphere), HARDI (dense unit sphere), and MS-HARDI (multiple shells).

where

$$\mathbf{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{2.10}$$

is the symmetric 3×3 co-variance matrix which defines the overall shape and orientation of the Gaussian distribution and has come to be known as the diffusion tensor (DT). Then, given the Fourier relationship with the diffusion signal, and the fact that the Fourier transform of a Gaussian is again a Gaussian, the EAP is given by

$$E(\mathbf{q}) = \exp(-4\pi^2 \tau \mathbf{q}^\top \mathbf{D}\mathbf{q}). \tag{2.11}$$

Since $\mathbf{D}$ is symmetric it contains 6 unknowns to estimate, requiring only 6 gradient directions in addition to $b = 0$. Furthermore, because we have a quadratic equation the domain of q-space can be restricted to the unit sphere where by $||\mathbf{q}|| = 1$. In practice, to deal with noise, anywhere from 10-60 DWIs may be acquired to estimate the DT with $b = 1000\ \mathrm{s/mm^2}$, making the acquisition process clinically advantageous.

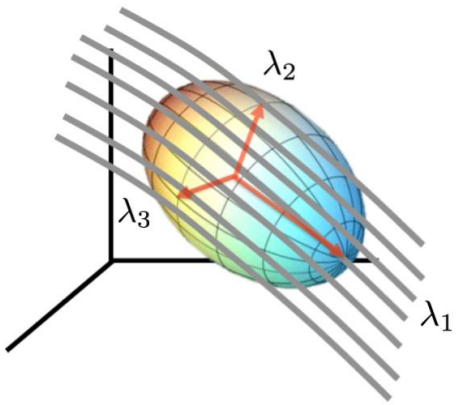


Figure 2.15: Diffusion tensor with eigenvalues λ_1, λ_2 and λ_3, describing the shape of the 3D Gaussian distribution with respect to the orientation of the underlying fiber population.

Researchers have derived many informative scalar features of the DT to better characterize its shape in relation to the degree of water diffusion. In particular, the eigenvalues of a covariance matrix, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ are equal to the lengths of the three oriented axes of the 3D Gaussian distribution, one major and two minor (see Figure 2.15). The most popular scalar feature, known as fractional anisotropy (FA), is a measure from 0 to 1 of the amount of diffusivity within a voxel as a function of

the eigenvalues:

$$FA = \sqrt{\frac{3}{2}}\sqrt{\frac{(\lambda_1 - \bar{\lambda})^2 + (\lambda_2 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}, \tag{2.12}$$

where $\bar{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$ is the mean of the eigenvalues, otherwise known as mean diffusivity (MD), and is also a common DTI feature (see Figure 2.16). Other popularly used DTI features include other functions of the DT eigenvalues and are used regularly in clinical studies to evaluate statistical differences of diffusivity and anatomical structure between normal and abnormal subjects.

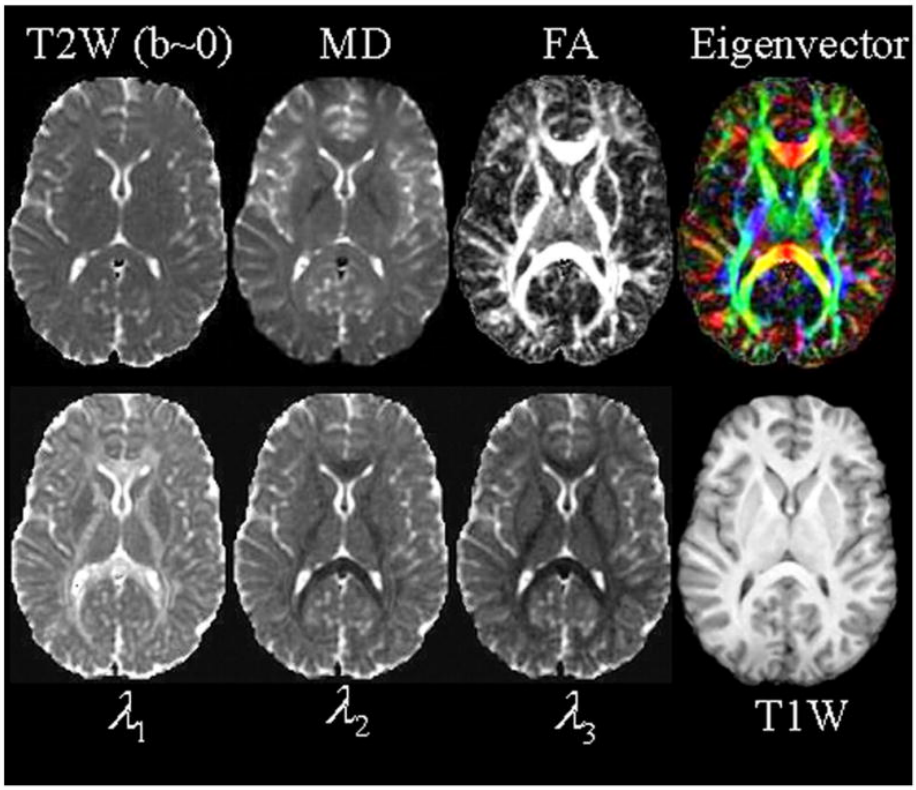


Figure 2.16: Diffusion information is extracted from DTI using features like FA, MD, principal eigenvector direction, and tensor eigenvalues in comparison to T1 and T2 weighted images. [Image adapted from [10].]

However, while DTI is celebrated for its simplicity and enjoys a host of clinically validated scalar features, a major drawback of its unimodal Gaussian model is the inability to represent more complex diffusion patterns that involve two or more crossing fibers in a single voxel or fanning, kissing, twisting or sheering of fiber tracts. To overcome these limitations, an alternative type of imaging protocol was developed which increases the angular resolution of spherical q-space samples allowing for higher order distributions to model the diffusion process.

2.1.3.3 High Angular Resolution Diffusion Imaging (HARDI)

High angular resolution diffusion imaging (HARDI) [31] was developed to overcome the model limitations of DTI by removing the single Gaussian model assumption, affording the ability to represent complex diffusion distributions of crossing fibers with multiple probabilistic peaks. Unlike DSI, HARDI still restricts q-space samples to the unit sphere, increasing the angular resolution from that of DTI anywhere from 60-200 or more gradient directions (see Figure 2.14). Instead of a Gaussian DT, HARDI aims to estimate a non-parametric spherical probability distribution, named the orientation distribution function (ODF), which indicates the probability of having a fiber tract along a given direction in a given voxel and can accurately model crossing fiber populations. The ODF is a very meaningful and interpretable feature of the EAP and can be calculated from it using DSI by integrating radially and projecting onto the unit sphere. Q-Ball Imaging (QBI) [42] was one early technique for calcu-

lating ODFs from spherical q-space samples (the q-ball), but later, a more correct derivation was developed by [43] which took into account the variation of integration over a solid constant angle. In particular, the ODF, p, is related to the EAP, P, by integrating radially over a solid constant angle:

$$p(\mathbf{r}) = \int_0^\infty P(R\mathbf{r})R^2 dR \tag{2.13}$$

where $\mathbf{r} = R\mathbf{r}$, $||\mathbf{r}|| = 1$. Using (2.13) and the definition of the EAP, the ODF can be written as a function of the normalized dMRI signal directly as:

$$p(\mathbf{r}) = \frac{1}{4\pi} + \frac{1}{16\pi^2}\text{FRT}(\nabla_b^2 \ln(-\ln E(\mathbf{q}))), \tag{2.14}$$

where $\text{FRT}(f(x)) = \int_{u \in C(x)} f(u)ds(u)$ with $C(x) = \{u \in \mathbb{S}^2 | u \cdot x = 0\}$ is the Funk-Radon Transform (FRT) and ∇_b^2 is the Laplace-Beltrami operator.

Spherical Harmonics. Since q-space samples for HARDI are confined to the unit sphere, the normalized diffusion signal can be written in terms of spherical basis. One well-studied choice is the spherical harmonic (SH) basis, which are continuous complex-valued functions on the unit sphere, defined as:

$$Y_l^m(\theta, \phi) = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta)e^{im\phi}, \quad l = 0, 1, 2, \ldots, -l \leq m \leq l, \tag{2.15}$$

where P_l^m is the associated Legendre polynomial of degree l and order m, $\theta \in [0, \pi]$,

and $\phi \in [0, 2\pi)$. Analogous to a Fourier basis on the sphere, the SH function elements can represent any spherical function. Figure 2.17 gives a visualization of the SH basis functions up to order $l = 4$. Now, since HARDI signals are real valued, it is more convenient to use the modified SH basis functions, defined as

$$Y_{l,m} = \begin{cases} \sqrt{2}Re(Y_l^{|m|}), & \text{if } -1 \leq m \leq 0, \\ Y_l^0, & \text{if } m = 0, \\ \sqrt{2}(-1)^{m+1}Im(Y_l^m), & \text{if } 0 \leq m \leq l, \end{cases} \tag{2.16}$$

and write $\ln(-\ln E(\mathbf{q})) = \sum_{l,m} c_{l,m} Y_{l,m}(\mathbf{q})$. The main advantage of this basis representation is that $Y_{l,m}$ are eigenfunctions of both the FRT and the Laplace-Beltrami operator $\text{FRT}(Y_{l,m}(\mathbf{q})) = 2\pi P_l(0) Y_{l,m}(\mathbf{r})$ and $\nabla_b^2(Y_{l,m}(\mathbf{q})) = -l(l+1)Y_{l,m}(\mathbf{q})$. Therefore, (2.14) simplifies to:

$$p(\mathbf{r}) = \frac{1}{4\pi} + \frac{1}{16\pi^2} \sum_{l,m} -2\pi P_l(0) l(l+1) c_{l,m} Y_{l,m}(\mathbf{r}) = \sum_{l,m} c_{l,m} \tilde{Y}_{l,m}(\mathbf{r}) \tag{2.17}$$

where $\tilde{Y}_{0,0} = \frac{1}{2\sqrt{\pi}} Y_{0,0}$ and $\tilde{Y}_{l,m} = \frac{1}{8\pi} P_l(0) l(l+1) Y_{l,m}$ for $l > 0$. In practice, when representing a signal, the SH functions are truncated to an order L, emitting $F = \frac{(L+1)(L+2)}{2}$ basis elements.

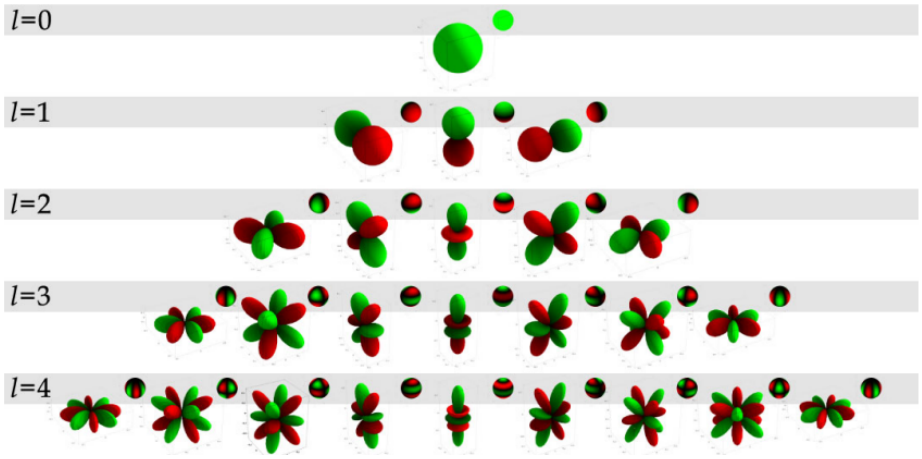


Figure 2.17: Visualization of spherical harmonic (SH) basis functions. For each order l, the rows of functions run from left to right with $m = -l$ to l.

Spherical Ridgelets/Wavelets. Another basis (or dictionary) used to represent HARDI signals and ODFs are the spherical ridgelets (SR) and spherical wavelets (SW), respectively [44]. While the SH basis is suitable for reconstructing any spherical bandlimited signals, the SR/SW pair is particularly advantageous for sparse representations of HARDI signals/ODFs. We will use the SR basis frequently in this thesis for sparse reconstruction of HARDI.

Derived using properties of spherical harmonics, the SR/SW pair are related by equation 2.14 and are known to provide sparse representations of HARDI signals/ODFs. As we will see in the coming thesis, sparsity is an important trait for sparse coding, de-noising and compressed sensing [45] for HARDI. Visually, the SW dictionary consists of a collection of single fiber ODFs rotated in 3D space which covers a large set of orientations (see Figure **??**). The shapes or scales of the ODF

dictionary elements range from nearly isotropic to highly anisotropic. Then, the SR are transformed from these oriented ODFs using the relation of equation (2.14).

Following [44, 45], the SR/SW bases are built using the SH property:

$$\sum_{l=-m}^{m} Y_{l,m}(\mathbf{u})Y_{l,m}(\mathbf{v}) = \frac{2l+1}{4\pi}P_l(\mathbf{u}\cdot\mathbf{v}), \tag{2.18}$$

for $\mathbf{u}, \mathbf{v} \in \mathbb{S}^2$ where P_l is the Legendre polynomial of degree l, for $l \geq 0$, defined as $P_l(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n$. It is noted that $P_l(\mathbf{u}\cdot\mathbf{v})$ is a rotation of $Y_{l,0}(\mathbf{u})$ about the vector $\mathbf{v}$ giving the orientational nature of the dictionary. Then, any symmetric square-integrable spherical function f can be written as:

$$f(\mathbf{u}) = \int_{\mathbb{S}^2} f(\mathbf{v}) \sum_{l=0,2,4,\ldots}^{\infty} \frac{2l+1}{4\pi}P_l(\mathbf{u}\cdot\mathbf{v})d\mathbf{v} = \int_{\mathbb{S}^2} f(\mathbf{v})\Theta_{\mathbf{v}}(\mathbf{u}), d\mathbf{v}, \tag{2.19}$$

where $\Theta_{\mathbf{v}}$ is a spherical convolution kernel. To account for multiple scales, the Gauss-Weierstrass scaling function $\chi_{\mathbf{v},j}$ at resolution $j \in \mathbb{N}$ is used, given by:

$$\chi_{\mathbf{v},j}(\mathbf{u}) = \sum_{l=0,2,4,\ldots}^{\infty} \frac{2l+1}{4\pi}\kappa_j(2^{-j}l)P_l(\mathbf{u}\cdot\mathbf{v}), \tag{2.20}$$

where $\kappa_\rho(x) = e^{-\rho x(x+1)}$. Then, for $\mathbf{r} \in \mathbb{S}^2$, the SW basis is constructed as the semi-discrete frame $\cup_{j=-1}^{\infty} \cup_{\mathbf{v}\in\mathbb{S}^2} \psi_{\mathbf{v},j}(\mathbf{r})$ with

$$\psi_{\mathbf{v},j}(\mathbf{r}) = \chi_{\mathbf{v},j+1}(\mathbf{r}) - \chi_{\mathbf{v},j}(\mathbf{r}), \tag{2.21}$$

for $j \geq 0$ and $\psi_{\mathbf{v},-1} \equiv \chi_{\mathbf{v},0}$. This can be rewritten using (2.20) in closed form as:

$$\psi_{\mathbf{v},j}(\mathbf{r}) = \sum_{l=2,4,\ldots}^{\infty} \frac{2l+1}{4\pi} \nu_{\rho,j}(l) P_l(\mathbf{r} \cdot \mathbf{v}) \tag{2.22}$$

with $\nu_{\rho,j}(l) = \kappa_\rho(2^{-j-1}l) - \kappa_\rho(2^{-j}l)$ for $j \geq 0$ and $\nu_{\rho,-1} \equiv \kappa_\rho(l)$. Then using (2.14), the corresponding SR basis in q-space is:

$$\gamma_{\mathbf{v},j}(\mathbf{q}) = \sum_{l=2,4,\ldots}^{\infty} \frac{4\pi(2l+1)}{-l(l+1)} \frac{\nu_{\rho,j}(l)}{\lambda_l} P_l(\mathbf{q} \cdot \mathbf{v}), \tag{2.23}$$

where λ_l and $-l(l+1)$ are the eigenvalues of the SH basis for the FRT and ∇_b^2, respectively. We will utilize the SR basis frequently in this thesis for sparse reconstruction of HARDI signals, while the companion SW is used to estimate ODFs.

HARDI Features. One popular scalar feature of the ODF, akin to the FA in DTI, is the generalized FA (GFA) [42] given by:

$$GFA = \sqrt{\frac{N \sum_{i=1}^{N} (p(\mathbf{r}_i)) - \bar{p})^2}{(N-1) \sum_{i=1}^{N} p(\mathbf{r}_i)^2}} \tag{2.24}$$

where $\{\mathbf{r}_i\}_{i=1}^{M}$ are the M discrete points where the ODF is evaluated on the unit sphere and $\bar{p} = \frac{1}{N} \sum_{i=1}^{M} p(\mathbf{r}_i)$ is the average. Like FA, the GFA computes a value between 0 and 1 indicating the degree of anisotropy of the fiber population. Many other features of the ODF have been proposed [46–49] including our own work [11] on rotation invariant features built from eigenvalues derived from the SH representation

of the ODF. Figure 2.18 illustrates the higher degree of information revealed from our proposed HARDI features over that of the GFA within a phantom dataset with known fiber populations in each voxel.

In our prior work [50], we proposed a novel method to unite feature extraction with registration and atlas building by registering scalar features that preserved the diffusion information extracted at every voxel to a novel feature atlas. Additionally our method provided an automatic way to select for the most informative features that drove the anatomical registration process. Feature selection is important machine learning challenge and useful for disease classification. In this vein, we compared the selected features of a healthy set of subjects with that of subjects who tested positive for beta-amyloid pathology, a predictor of Alzheimer's disease. We compared features extracted from various diffusion models and found that features proposed in our prior work [11] may be most distinguishable between these two populations.

2.1.3.4 Multi-Shell HARDI and Other Higher Order Models

Sampling q-space on the unit sphere, saves a significant amount of time over the dense Cartesian sampling of DSI. To provide an intermediate protocol, HARDI was extended to sample along multiple shells of varying b-values in order to estimate the EAP without fully sampling q-space. This has been called Multi-Shell HARDI (MS-HARDI). To calculate the EAP from MS-HARDI, a common formulation is to

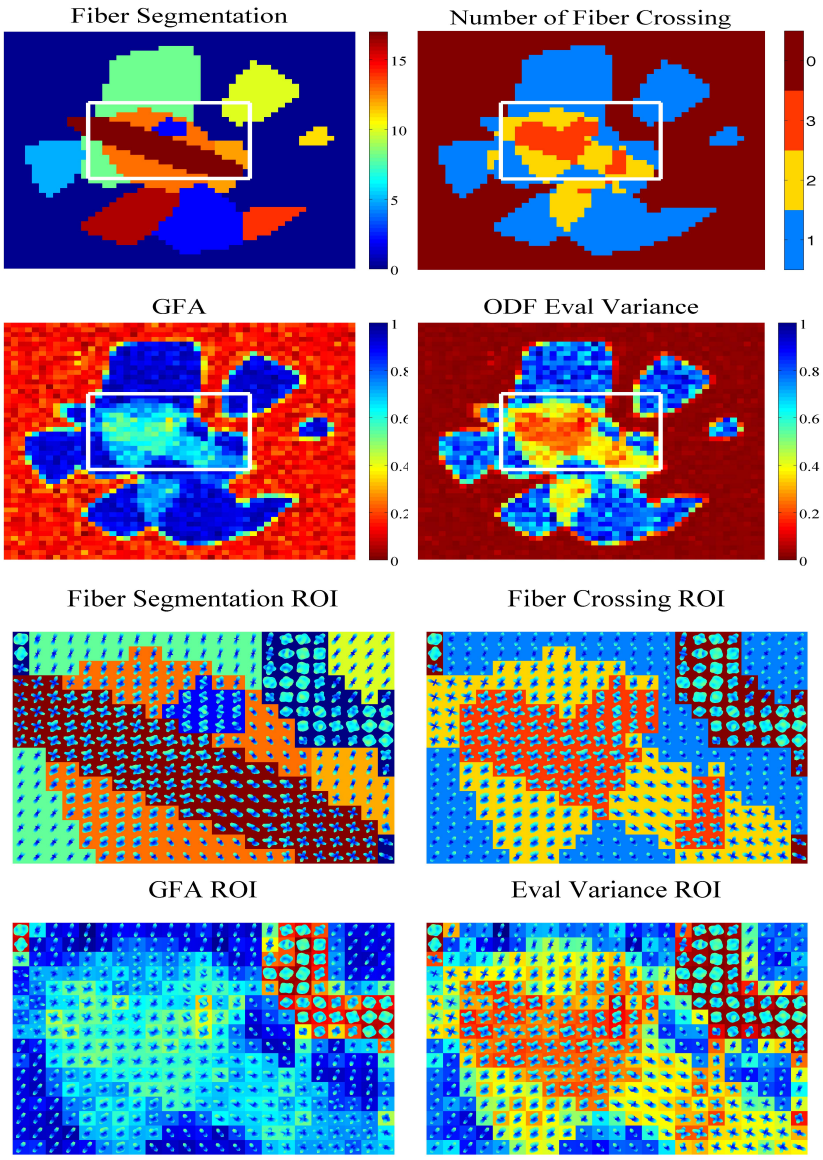


Figure 2.18: Analysis of HARDI features extracted from the ISBI 2013 HARDI Phantom dataset. First Row: The left image is the ground truth fiber segmentation of a slice of the phantom dataset, where the rectangle highlights an ROI with an intricate region of crossing fibers. The right image is a count of the number of fibers that cross in a given voxel, ranging from 0 to 3. Second Row: GFA and eigenvalue variance of the phantom slice. We notice here the striking similarity between the plot of crossing fibers and the eigenvalue variance whereas the GFA is unable to reveal this information. Third/Fourth Row: Close up of the ROI with ODFs. [Image adapted from our prior work [11].]

incorporate a radial function to the SH representation [51]:

$$P(\mathbf{r}) = \sum_{l,m,n} c_{lmn}\rho_n(R)Y_{lm}(\mathbf{r}) \tag{2.25}$$

where the radial ρ_n are known as the Gauss-Laguerre basis functions:

$$\rho_n(R) = \big(\frac{2n!}{\xi^{3/2}\Gamma(n+3/2)}\big)^{1/2}\exp(\frac{-R^2}{2\xi})L_n^{1/2}(\frac{R^2}{\xi}), \tag{2.26}$$

where ξ is a scale factor and $L_n^{1/2}$ is the generalized Laguerre polynomial of order $n \in \mathcal{N}$, and Γ is the Gamma function. The combined $\rho_n(R)Y_{lm}(\mathbf{r})$ elements are called the Spherical Polar Fourier basis.

In addition to single- and multi-shell HARDI, there are a vast number of bases, models and representations for reconstructing dMRI signals and estimating PDFs of fiber orientation that are beyond the scope of this thesis. Some of these include spherical deconvolution [52] which aims to estimate a *fiber* ODF that can better resolve crossing fibers of small angular separation, higher-order tensors [53] which act as a generalization to the 2nd order DT, multi-tensor methods [31] which include multiple Gaussian fiber populations aim to model the microstructure of fiber tracts with multiple diffusion compartments.

With an overview of the many diffusion models available in the literature, the goal now is to estimate the distribution of fiber orientation from measured dMRI signals. In the next section, I will outline the task of signal reconstruction and the estimation

of ODFs specifically for the case of single-shell HARDI acquisition.

2.1.4 Signal Reconstruction and ODF Estimation for HARDI

As outlined in the previous section, a single-shell HARDI signal with G gradient directions $\{\mathbf{q}_j\}_{j=1}^{G} \in \mathbb{S}^2$ can be modelled as a linear combination of SH basis functions and the coefficients can be used to estimate the ODF. Specifically, the HARDI signal in a given voxel $s = [s(\mathbf{q}_1), \ldots, s(\mathbf{q}_G)]^\top \in \mathbb{R}^G$, can be written as:

$$s = Bc + \epsilon \tag{2.27}$$

where $B = [Y_1(\mathbf{q}), \ldots, Y_F(\mathbf{q})] \in \mathbb{R}^{G \times F}$ is the matrix of modified SH basis functions of degree L (recall $F = \frac{(L+1)(L+2)}{2}$ atoms as per (2.16)) evaluated at points $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_G]^\top$, $c \in \mathbb{R}^R$ is the set of SH coefficients which parameterize s, and ϵ approximates unmodeled observation errors. A common choice for order L is 4, with $F = 15$. To find c we solve the least-square problem

$$\min_c \frac{1}{2}||Bc - s||_2^2 \tag{2.28}$$

Then, the reconstructed signal is $s^* = Bc^*$ and by (2.17) the ODF is constructed as $p^* = \tilde{B}c^*$, where $\tilde{B} = [\tilde{Y}_1(\mathbf{r}), \ldots, \tilde{Y}_F(\mathbf{r})] \in \mathbb{R}^{M \times F}$ is evaluated at M selected points

$\{\mathbf{r}_i\}_{i=1}^{M} \in \mathbb{S}^2$ with $\mathbf{r} = [\mathbf{r}_1, \ldots, \mathbf{r}_M]^\top$.

However, ODFs are PDFs on the unit sphere, and therefore should be non-negative and sum to 1, conditions which are often violated due to noisy measurements. While ensuring the distribution sums to 1 is done by rescaling after estimation, enforcing non-negativity has been addressed by adding constraints on the SH coefficients in (2.28) as [32]:

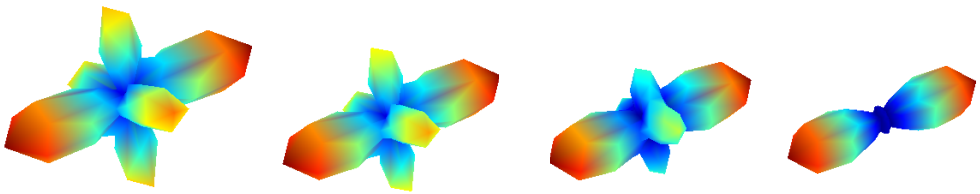
$$\min_c \frac{1}{2}||Bc - s||_2^2 \;\text{ s.t. }\; \tilde{B}c \geq 0. \tag{2.29}$$

However, this constraint ensures that the estimated ODF be non-negative only at the M selected points $\{r_i\}$. In our prior work [12, 54], we improve upon this result by enforcing that the ODF be non-negative everywhere on the continuous domain:

$$\min_c \frac{1}{2}||Bc - s||_2^2 \;\text{ s.t. }\; \tilde{B}(\mathbf{r})c \geq 0 \;\; \forall\, \mathbf{r} \in \mathbb{S}^2, \tag{2.30}$$

where $\tilde{B}(\mathbf{r}) = [\tilde{Y}_1(\mathbf{r}), \ldots, \tilde{Y}_F(\mathbf{r})]$. This becomes a difficult problem because of the infinite number of constraints over the continuous domain. In [12], we enforce non-negativity on a set of eigenvalues extracted from a transformation of the SH coefficients and solve a new positive semi-definite optimization (see Figure 2.19). This method was improved in [54], by realizing that enforcing non-negativity at an infinite number of points is equivalent to enforcing non-negativity at a single active constraint at the minimum.

Enforcing non-negativity of ODFs is one such constraint used in signal reconstruc-



(a) LS (b) DN (c) CN (d) True

Figure 2.19: Non-negative ODF estimation using (a) Least Squares (LS) (2.28), (b) Discrete Non-negativity (DN) (2.29), and (c) Continuous Non-negativity (CN) (2.30) compared to (d) the ground truth ODF. We compare the reconstructions of the three methods for a single fiber ODF with SNR 5 dB. Our method CN provides a more accurate reconstruction by reducing negative lobes resulting from noisy data. [Image adapted from our prior work [12].]

tion. In general, there are many other properties of interest used in signal processing, like enforcing spatial regularity, ODF smoothing, de-noising, interpolation, and super-resolution which can be solved by constraints in signal reconstruction. Then once signals are reconstructed and ODFs are estimated accurately, they are used to reconstruct fiber tracts in tractography. However, before dMRI can be used for tractography and disease applications, it must be acquired at a clinically acceptable rate. In the next section, we introduce methods for accelerating dMRI acquisition.

2.1.5 Acceleration of HARDI Acquisition and Reconstruction

While dMRI protocols like DSI, HARDI and MS-HARDI provide more accurate models of diffusion than DTI, the complexity of their models usually implies a much higher number of q-space samples resulting in long, clinically unsuitable scan times

relative to DTI. A major ongoing research goal has been to reduce acquisition times of these advanced imaging techniques, while maintaining accurate estimations of diffusion. To date, two main paths to accelerating image acquisition have been explored. The first one is from a hardware perspective and involves new methods in parallel and multi-slice imaging to acquire multiple signals from a single subject simultaneously. The second one is from a signal processing processing perspective, using a paradigm known as compressed sensing, which involves sparse representations of the data for which the full number of signal measurements is no longer necessary to reconstruct a full image. Ideally, the two areas may be optimally integrated to further accelerate acquisition and this has been an ongoing effort as well. They are summarized briefly below.

2.1.5.1 Parallel/Multi-Slice Imaging

The goal of parallel imaging is to accelerate MR signal acquisition by subsampling along the phase-encoding direction in k-space. However, sampling at sub-Nyquist levels will result in a reduced field of view (FOV) and aliasing artifacts in the form of double overlapping images. The idea behind parallel imaging is instead of using a single receiver coil over the entire slice, to utilize an array of receiver coils each positioned to acquire smaller FOV images along a reduced set of phase-encoding k-space measurements (see Figure 2.20). Therefore a full FOV image can be imaged in parallel by the combination of each local receiver coil and the full MR image can be

acquired in an accelerated rate proportional to the number of coils in the array.

The challenge then is to recombine the individual sensor data and remove their spatial aliasing. A number of algorithms have been proposed for accelerated reconstruction like sensitivity encoding (SENSE) [55] and generalized autocalibrating partially parallel acquisitions (GRAPPA) [56]. SENSE uses the locations of the local coils to solve a number of linear programs by constraining overlapping pixel data in the reconstructed spatial domain. On the other hand, GRAPPA aims to recover the missing k-space values by the overlapping information in neighboring coils and then reconstruct a full MR image. While parallel imaging refers to multiple coils sensing simultaneously in the phase-encoding k-space dimension, simultaneous multi-slice (SMS) imaging refers to sensing multiple slices of an MR volume in parallel. Both parallel and SMS imaging methods have been extended to the domain of dMRI [57] and even further combined with the framework of compressed sensing [58] which we summarize next.

2.1.5.2 Compressed Sensing

Compressed Sensing (CS) is a class of mathematical results and algorithms that exploits sparse representations of signals, discovered through sparse coding, to obtain extremely accurate reconstructions at sub-Nyquist rates [59]. The main ingredients of the CS framework are an appropriately chosen sampling scheme and an underlying "sparse" representation of the data. The key idea is that, the sparser the representa-

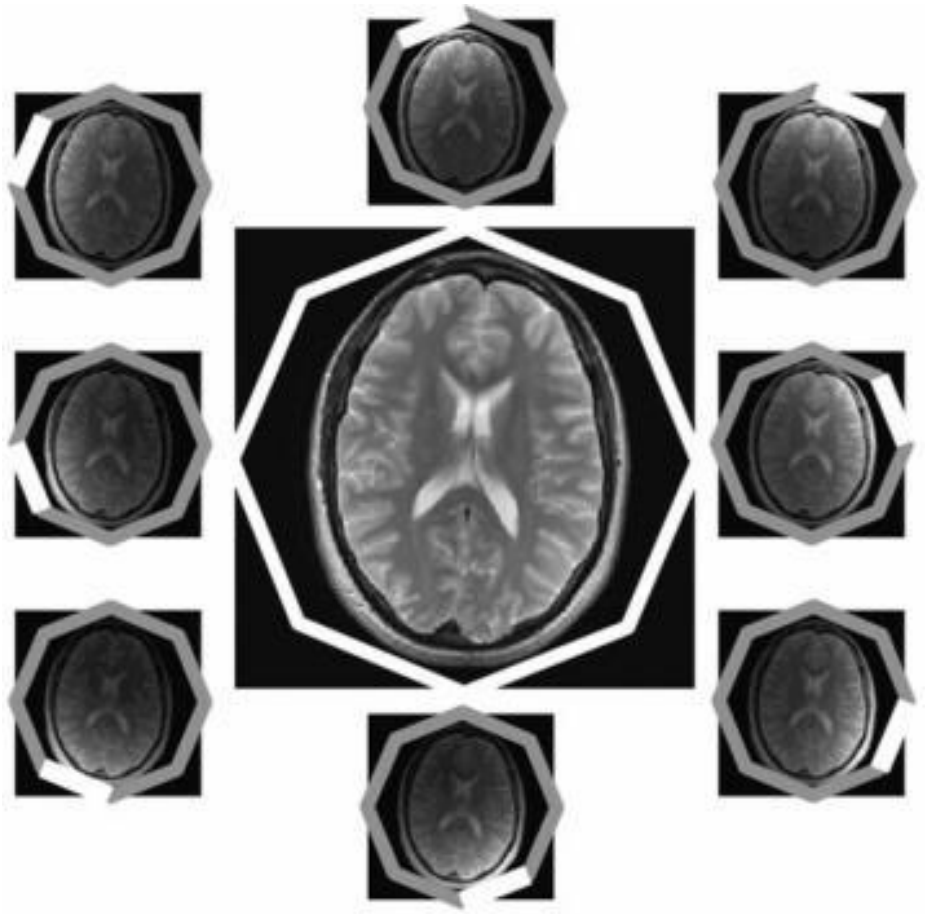


Figure 2.20: Example acquisition with 8 receiver coils positioned around the subject. Parallel imaging accelerates acquisition by imaging subsets of an entire brain image in parallel and reconstructing the whole image using post-processing algorithms based on the known locations of each coil. [Image adapted from [13].]

tion, the fewer the samples needed to reconstruct the full signal with high accuracy.

CS has been classically applied to MRI [60] by subsampling in the native k-space (k-CS) while applying sparsifying transforms in the spatial image domain like wavelets and total-variation (TV). For dMRI, diffusion signals are measured along different angular gradient directions in q-space for every point in k-space. Thus, to reduce the number of diffusion measurements, many methods [61] have exploited

sparse representations in the angular domain by applying CS in q-space (q-CS). To further accelerate dMRI, more recent methods [62] combine aspects of k-CS and q-CS by subsampling jointly in (k, q)-space ((k, q)-CS).

One of the core contributions of this thesis is new advances in CS for dMRI and so in the following section, we will review the ingredients of CS in the larger context of sparse reconstruction for general signals.

2.2 Principles of Sparse Reconstruction

The main contributions of this thesis are to advance the fields of sparse coding, compressed sensing, and dictionary learning for dMRI signal processing and acceleration. In this section, we introduce the basic concepts of sparse coding, compressed sensing, and dictionary learning for general signals, to be used as a reference for the remainder of the thesis. We will also include an introduction to convolutional methods applied to sparse coding and dictionary learning.

2.2.1 Sparse Coding

In signal processing, a well studied problem is that of reconstructing a signal from a set of noisy measurements by finding a representation of the signal in a chosen domain for which one can more easily process and analyze the data. This is a well-studied problem that has been widely used in applications in machine learning [63]

such as de-noising [64], super-resolution [65] and neural networks [66] and can be used to combat over-fitting of signal reconstruction.

2.2.1.1 Formulation

In the most general setting of representation theory one would like to represent a signal $s \in \mathbb{R}^N$ in terms of a dictionary $D \in \mathbb{R}^{N \times N_D}$ with N_D dictionary atoms as

$$s = Da + \epsilon, \tag{2.31}$$

where the coefficient vector $a \in \mathbb{R}^{N_D}$ is the representation of s in terms of dictionary D and ϵ represents unmodeled errors or deviations from the model. The procedure of finding the coefficients a becomes an optimization problem:

$$\min_a \ell(s, Da), \tag{2.32}$$

where ℓ is defined as some "loss" function which penalizes some type of difference between the measured signal s and the reconstruction $\hat{s} = Da$. The most common choice of ℓ, is the squared norm of the difference:

$$\ell(s, Da) = \frac{1}{2}||s - Da||_2^2 \tag{2.33}$$

where $||\cdot||_2$ is the usual L_2 norm defined by $||x||_2 = \sqrt{\sum_i^N |x_i|^2}$ for vector $x \in \mathbb{R}^N$. Depending on the size of D, an exact solution can be solved using least squares. But to avoid overfitting the noise in the signal, including additional regularization on the solution space of a may be necessary to produce more structured representations.

One possible regularization strategy is to require the solution of (2.32) to be sparse, i.e. the vector a contains very few non-zero elements. The most notable application of sparse coding in our context is that of compressed sensing, by which the level of sparsity, i.e. the number of non-zero elements, is closely related to the amount of subsampling of the signal that can be achieved. Formally, the sparse coding problem can be written as:

$$\min_a \ g(a) \ \text{ s.t. } \ \frac{1}{2}||s - Da||_2^2 \leq \epsilon, \tag{2.34}$$

where g is a regularizer that promotes sparsity. In the next section we will preview some of the many algorithms designed to solve this general sparse coding problem, depending on the choice of regularizer g.

2.2.1.2 Algorithms

The most fundamental choice of regularizer g to measure sparsity is the L_0 semi-norm, $g(a) = ||a||_0$, which simply counts the number of non-zero entries of vector a. Solving (2.34) with the L_0 semi-norm however is intractable and so a family of greedy algorithms have been proposed to approximate a solution. For example, the popular

Orthogonal Matching Pursuit (OMP) algorithm [67] selects the K dictionary atoms most correlated with the signal one by one, removing the residual of the currently selected atom, orthogonalizing the coefficients, and then finding the next most correlated. The OMP algorithm is stated in Algorithm 1, where at iteration k, $\mathcal{I}^k$ is the set of indices for chosen dictionary atoms, and r_k is the residual initialized as the signal itself. From the orthogonality property of the residual at each iteration, the algorithm is guaranteed to find a next most correlated atom and converge in a finite number of iterations.

Algorithm 1 Orthogonal Matching Pursuit (OMP)

Choose: K, ϵ.
Initialize: $k = 1, \mathcal{I}^0 = \emptyset, r_0 = s$.
while $k \leq K$ and error $> \epsilon$ **do**
  $i^k = \arg\max_i |\langle r_k, D \rangle|$;
  $\mathcal{I}^k = [\mathcal{I}^{k-1}, i^k]$;
  $a_k = \arg\min_a ||s - D_{\mathcal{I}^k} a||_2^2$;
  $r_k = s - D_{\mathcal{I}^k} a_k$;
  $k \leftarrow k + 1$;
end while

To avoid the use of greedy algorithms, another approach is to relax the problem by using the L_1 norm, $g(a) = ||a||_1$, which is known to promote sparse solutions. Then (2.34) can be reformulated using a relaxation term as

$$\min_a \ \frac{1}{2}||s - Da||_2^2 + \lambda||a||_1, \tag{2.35}$$

and solved using convex optimization techniques. Problem (2.35) is commonly known as the Least Absolute Shrinkage and Selection Operator or LASSO problem, due to

the L_1-norm's ability to shrink the absolute value of some of the coefficients and reduce others to 0 [68].

Gradient descent is a common method for solving convex optimization problems which do not have closed form solutions of the optimality equations. However, because the L_1 norm is non-differentiable an alternative strategy is to combine the gradient with the proximal operator of the non-smooth part. The definition of the proximal operator of a non-differentiable function $g(a)$ is:

$$\text{prox}_{\lambda g(\cdot)}(a) = \arg\min_x \frac{1}{2\lambda}||x-a||_2^2 + g(x), \tag{2.36}$$

When $g(a) = ||a||_1$, the proximal operator has a closed-form expression given by the shrinkage or soft-thresholding operator:

$$\text{prox}_{\lambda\|\cdot\|_1}(a) \equiv \text{shrink}_\lambda(a) = \text{sign}(a) \cdot \max(|a| - \lambda, 0), \tag{2.37}$$

where the max is taken element-wise over the vector a. Algorithms that utilize this result include the Fast Iterative Shrinkage Thresholding Algorithm (FISTA), stated in Algorithm 2, which performs a proximal gradient descent on (2.35). Letting $f(a) = \frac{1}{2}||s - Da||_2^2$, we take a step in the gradient direction $\nabla f(a) = D^\top s - D^\top Da$ and then take the shrinkage operator. FISTA outperforms its ISTA precursor by including Nesterov acceleration (end of Algorithm 2).

Another popularly used algorithm is known as the Alternating Direction Method

Algorithm 2 Fast Iterative Shrinkage Thresholding Algorithm (FISTA)

Choose: L, λ, ϵ.
Initialize: $k = 1, z_1 = a_0 = 0$, $n_1 = 1$.
while error $> \epsilon$ **do**
  $a_k = \text{shrink}_{\lambda/L}(z_k - \nabla f(z_k)/L)$;
  $n_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_k^2})$;
  $z_{k+1} = a_{k+1} + \frac{n_k - 1}{n_{k+1}}(a_{k+1} - a_k)$;
  6: $k \leftarrow k + 1$;
end while

of Multipliers (ADMM) which is a universal approach to solving convex optimization problems that may involve one differentiable and one non-differentiable function, as is the case of LASSO, by splitting the problem into two subproblems and alternatingly updating each variable. More specifically, with a change of variable, (2.35) is equivalent to

$$\min_{a,v} \ \frac{1}{2}||s - Da||_2^2 + \lambda||v||_1 \ \text{ s.t. } v = a \tag{2.38}$$

Then incorporating the constraint through the augmented Lagrangian problem with Lagrange multiplier τ and additional relaxation with constant μ the problem can be posed as

$$\min_{a,v} \max_{\tau} \ \{L_\mu(a, v, \tau) = \ \frac{1}{2}||s - Da||_2^2 + \lambda||v||_1 + \langle \tau, a - v \rangle + \frac{\mu}{2}||a - v||_2^2\}. \tag{2.39}$$

ADMM alternates optimization over a, v, and τ until convergence with the updates described in Algorithm 3.

Algorithm 3 Alternating Direction Method of Multipliers (ADMM)

Choose: μ, λ, ϵ.
Initialize: $k = 0, v_0 = 0, \tau_0 = 0$.
while error $> \epsilon$ **do**
  $a_{k+1} = \arg\min_a L_\mu(a, v_k, \tau_k)$;
  $v_{k+1} = \arg\min_v L_\mu(a_{k+1}, v, \tau_k)$;
  $\tau_{k+1} = \tau_k + a_{k+1} - v_{k+1}$;
  $k \leftarrow k + 1$;
end while

2.2.2 Compressed Sensing

Compressed Sensing (CS) is a class of mathematical and signal processing results aimed at recovering a signal from a compressed set of measurements. Classical sampling theory and algorithms were designed for the class of band-limited signals. In this case, the Shannon-Nyquist theorem states that the minimum sampling rate of a band-limited signal is twice the highest frequency component of the signal, otherwise, sampling below this Shannon-Nyquist rate would induce aliasing in the reconstruction.

On the other hand, CS theory is designed for signals (1) that admit a sparse representation with respect to some dictionary and (2) for which the choice of dictionary and sampling pattern provide adequate conditions for recovering the original signal. Under these conditions, CS theory predicts that one can sample below the Shannon-Nyquist rate and still recover the full signal.

2.2.2.1 Formulation

To motivate these two themes, the problem we consider is recovering a signal $s \in \mathbb{R}^N$ of full dimension N, from a subset of noisy samples $\hat{s} \in \mathbb{R}^m$, with $m << N$, as

$$\hat{s} = \mathcal{U}s \tag{2.40}$$

where $\mathcal{U} \in \mathbb{R}^{m \times N}$ is a sensing or under-sampling matrix applied to s. Recovering s from (2.40) is ill-posed because there are an infinite number of solutions to this under-determined problem. By imposing a structure on the signal s, the space of solutions can be constrained. The structure that is assumed is that s is sparse with respect to some dictionary, i.e. s can be written as

$$s = Da \;\; \text{s.t.} \; ||a||_0 \leq K \tag{2.41}$$

for some non-zero sparsity level K. Combining (2.40) and (2.41), the samples are a function of the underlying sparse code a as:

$$\hat{s} = \mathcal{U}Da = \Phi a. \tag{2.42}$$

Then given $\hat{s}$, the goal is to minimize the number of non-zero coefficients of a such that $\hat{s} = \mathcal{U}Da$, known as the P_0 problem:

$$(P_0) : \min_a \ ||a||_0 \ \text{ s.t. } \ \hat{s} = \mathcal{U}Da \tag{2.43}$$

As we saw in the case of sparse coding in the previous section, because of the L_0 semi-norm, P_0 is a combinatorial NP-hard problem. Methods such as OMP can find an approximate solution and work well in practice. Alternatively, one can relax the L_0 semi-norm to the L_1 norm and solve the following convex Basis Pursuit problem [69]:

$$(P_1) : \min_a \ ||a||_1 \ \text{ s.t. } \ \hat{s} = \mathcal{U}Da. \tag{2.44}$$

The above formulations of P_0 and P_1 with exact constraints may not be adequate for real applications due to noise and other sampling artifacts. Letting ξ be a vector that captures the deviations from the ideal model, a noisy version of the exact constraints can be written as:

$$\hat{s} = \mathcal{U}Da + \xi. \tag{2.45}$$

Then, the problems to consider instead are:

$$(P_0^\epsilon) : \min_a \ ||a||_0 \ \text{ s.t. } \ ||\hat{s} - \mathcal{U}Da||_2^2 \leq \epsilon \tag{2.46}$$

and

$$(P_1^\epsilon) : \min_a \ ||a||_1 \ \text{ s.t. } \ ||\hat{s} - \mathcal{U}Da||_2^2 \leq \epsilon. \tag{2.47}$$

where the constant $\epsilon > 0$ is an error threshold. The former (2.46) can again be approximated using OMP and the latter (2.47) can be solved using Basis Pursuit Denoising. As discussed in the previous section, problem P_1^ϵ can also be rewritten with Lagrange multipliers and solved using LASSO algorithms.

2.2.2.2 Recovery Conditions

The algorithms we have discussed work well in practice to a find a sparse solution a which can synthesize a full resolution signal, but theoretically it is useful to know under what conditions we can guarantee the recovery of the true set of coefficients that generated our data. A key notion important for guaranteeing signal recovery is known as the Restricted Isometry Property (RIP) which attempts to measure how close a matrix Υ is to being an isometry, i.e. obeying $||\Upsilon a||_2^2 = ||a||_2^2$, when restricted to sparse vectors.

Definition 1. A matrix Υ satisfies the RIP property if:

$$(1 - \delta_K)||a||_2^2 \leq ||\Upsilon a||_2^2 \leq (1 + \delta_K)||a||_2^2, \tag{2.48}$$

for all vectors a with $\|a\|_0 \leq K$, where δ_K, the isometry constant of Υ, is taken to be the smallest of the constants such that the RIP property holds.

Satisfying the RIP property guarantees that we can distinguish between multiple K-sparse solutions and recover the unique solution to our problem.

Theorem 2. *If $\Upsilon \doteq \mathcal{U}D$ satisfies the RIP property with $\delta_{2K} < \sqrt{2} - 1$, the solution a^* to problem P_1^ϵ (2.47) obeys*

$$||a^* - a||_2 \leq C_0 K^{-1/2} ||a_K - a||_1 + C_1 \epsilon. \tag{2.49}$$

where a_K contains the K largest magnitude elements of a and zero otherwise, and C_0 and C_1 are constants.

Several classes of random subsampling matrices like Gaussian, Bernoulli,... have been shown to verify RIP with overwhelming probability [70]. However, with fixed deterministic sampling schemes, the RIP property is very difficult to check in practice as it requires testing all possible K-sparse subspaces, which is combinatorial.

Another condition for signal recovery, known as mutual coherence, was also proposed to provide a more practical criterion. In that case, considering a fixed orthonormal measurement matrix $\Phi \in \mathbb{R}^{N \times N}$ (from which we will extract a subset of rows to obtain the subsampling matrix $\mathcal{U}$), one defines:

Definition 3. The mutual coherence, $\mu(\Phi, D)$, of the sensing matrix Φ and dictionary D is the maximum absolute value of the inner products between all the columns of Φ and D, i.e.

$$\mu(\Phi, D) = \max_{1 \leq i \leq N, 1 \leq j \leq N_D} \langle \Phi_i, D_j \rangle. \tag{2.50}$$

This quantity measures the degree of similarity between the measurement system and the dictionary. When D is itself a $N \times N$ orthonormal matrix (e.g Fourier basis, Haar wavelet basis...), we have $\mu \in [1/\sqrt{N}, 1]$. Then low coherence $\mu \approx 1/\sqrt{N}$ essentially means that the measurements in Φ are very spread out on the domain of D whereas high coherence $\mu \approx 1$ implies maximal concentration along certain components of D. Coherence and sparsity of the underlying signal are then related to the number of measurements M needed to recover s by the following result:

Theorem 4. *Consider a subset $T \subset \{1, \ldots, N\}$ with $|T| = M$ chosen uniformly at random and the matrix $\mathcal{U}$ as the restriction of Φ to the column indexed by T. Then, with*

$$M \geq CK\sqrt{N}t\mu(\Phi, D)\log(tK\log N)\log^2 K \tag{2.51}$$

where C is a constant, and with probability exceeding $1 - e^{-t}$ over the choice of T, the matrix UD satisfies the RIP condition with $\delta_{2K} < \sqrt{2} - 1$.

Therefore, combining both Theorems 2 and 4, the more incoherent Φ and D are (smaller μ), and the sparser the representation (smaller K), the fewer measurements M are needed to recover s from its subsamples. Adequate sparsifying dictionaries for the signals at hand may be given by classical bases like wavelet. While general dictionaries may produce sparse representations in certain signals, it is also possible to optimize that choice by learning the dictionary directly from a training set of signals. This is an actively researched field known as *dictionary learning* that we summarize

in the following section and develop in Chapter 5 for the specific case of HARDI.

Note however that RIP or low coherence assumptions are usually not well-adapted to signals represented in overcomplete (instead of orthogonal) dictionaries D since such dictionaries may typically have highly correlated atoms. In such cases, more adequate notions and recovery guarantees can be considered, which will be discussed more in detail in the next section.

2.2.2.3 Recovery Conditions with Overcomplete Dictionaries

For overcomplete/redundant dictionaries, which have more columns then rows (commonly used for sparse coding), reconstruction guarantees derived from the notion of incoherence as in Section 2.2.2 are usually weak since redundant dictionaries tend to be highly coherent in practice. In [71], it is argued that incoherence is in fact not a necessary restriction for accurate reconstruction of signals in such cases. Instead, the authors proposed an alternative notion that extends the Restricted Isometry Property (RIP) presented in Section 2.2.2, which they coined D-RIP for RIP adapted to a certain dictionary D. In this framework, for a given dictionary D (and associated transform D^*), it is defined as follows:

Definition 5. A sensing matrix $\mathcal{U}$ is said to satisfy the D-RIP property with constant δ_K if

$$(1 - \delta_K)\|v\|_2^2 \leq \|\mathcal{U}v\|_2^2 \leq (1 + \delta_K)\|v\|_2^2 \tag{2.52}$$

for all vector v belonging to the reunion of all subspaces spanned by K columns extracted from D.

Then, given a measurement vector $y = \mathcal{U}s + z$ where s is the ideal signal and z a noise vector with $\|z\|_2 \leq \epsilon$, one introduces the usual L_1-analysis basis pursuit problem:

$$\tilde{s} = \arg\min_{s} \ \|D^* s\|_1 \ \text{ subject to } \ \|\mathcal{U}s - y\|_2 \leq \epsilon \tag{2.53}$$

and the following important result of stable reconstruction is shown in [71]:

Theorem 6. *If D is a tight frame and $\mathcal{U}$ a sensing matrix satisfying the D-RIP property with constant $\delta_{2K} < 0.08$ then the solution $\tilde{s}$ to (2.53) satisfies:*

$$\|\tilde{s} - s\|_2 \leq C_0 \epsilon + C_1 \frac{\|D^* s - (D^* s)_K\|_1}{\sqrt{K}}$$

for some constants C_0 and C_1, where $(D^* s)_K$ denotes the K largest coefficients of the decomposition $D^* s$.

In particular, in the noiseless scenario and if the signal s is K-sparse, then the previous theorem shows that it can be recovered exactly from the measurements by solving (2.53). In the general case, we see that the reconstruction error given by the right hand term will be smaller if the signal is more compactly decomposed by D, provided that the sensing matrix satisfies the adequate D-RIP property. We will revisit the implications of D-RIP for overcomplete dictionaries in Chapter 4.4.

2.2.3 Dictionary Learning

In sparse coding, the sparsity of a is dependent on the choice of dictionary D and for different types of signals there may be an array of known dictionaries that produce sparse representations (e.g. Wavelets for natural images). However, prescribing a known dictionary for a new signal or data type may lead to suboptimal sparsification (amount of sparsity of the representation) with respect to a specific signal of interest. To overcome this limitation, the idea of learning adapted dictionaries from training examples of typical signals of interest is known as dictionary learning.

Given a training set of T signals $\{s_t\}_{t=1}^{T}$ that resemble the signal of interest, dictionary learning amounts to solving (2.34) jointly over D and each coefficient vector a_t with added constraints on the dictionary atoms $\{D_k\}_{k=1}^{N_D}$ to keep them from growing in magnitude:

$$\min_{\{a_t\},D} \sum_{t=1}^{T} g(a_t) \quad \text{s.t.} \quad \frac{1}{2}\sum_{t=1}^{T} ||s_t - Da_t||_2^2 \leq \epsilon, \quad ||D_k||_2^2 \leq 1 \quad \forall\ k. \tag{2.54}$$

There have been many proposed algorithms to solve this dictionary learning problem, that alternate between solving for $\{a_t\}$ while D is fixed (sparse coding update), and solving for D while $\{a_t\}$ are fixed (dictionary learning update). One of the earliest algorithms is the Method of Optimal Directions (MOD) [72] which alternates between sparse coding and using the analytical least-squares pseudoinverse to solve for D.

Perhaps the most well-known dictionary learning algorithm is called K-SVD [73]

which solves the dictionary update for fixed sparse codes $\{a_t\}$, by solving a singular value decomposition (SVD) for each dictionary atom. More specifically, the dictionary update of K-SVD decomposes the multiplication DW (where $W = [a_1, \ldots, a_T]$) as a sum of T rank-1 matrices $d_t w_\top^t$, where $w_\top^t$ denotes the t^{th} row of W. Letting $Y = [s_1, \ldots, s_T]$, each atom d_t can be solved in succession by noting:

$$||Y - DW||_F^2 = ||(Y - \sum_{t \neq j} d_t w_\top^t) - d_j w_\top^j||_F^2 = ||E_j - d_j w_\top^j||_F^2. \tag{2.55}$$

Since $w_\top^j$ is sparse, the authors reduce the size by keeping only the nonzero elements which correspond to the K training signals in Y that use d_j. Defining E_j^K as the reduced matrix E_j with the K columns corresponding to the nonzero entries of $w_\top^j$ removed, d_j can be solved for by taking the SVD of E_j^K, where d_j is the left singular vector. Once a dictionary is learned, one can use this in (2.34) to sparsely represent a new (test) signal of interest.

Dictionary learning has been shown to outperform analytic dictionaries in terms of sparse reconstruction and applications like de-noising and image processing for natural and medical images. Because learning dictionaries directly from signals of interest can produce sparser codes, this has also been shown to be useful for reducing measurements via compressed sensing. One major drawback of the dictionary learning problem however, is that the joint optimization of the sparse codes and the dictionary is non-convex, hence existing alternating minimization algorithms may get stuck at

local minima. A major challenge in the field of dictionary learning is deriving methods which can guarantee globally optimal solutions and addressing this limitation will be one of the main contributions in this thesis.

One popular domain of our interest for learning dictionaries is for image data. However, images may be somewhat large in size, especially as the resolutions of modern cameras and medical imaging modalities continue to increase, and so learning global dictionaries becomes computationally difficult. In addition, natural and medical images exhibit diverse local textures and patterns, making local dictionary learning advantageous for some applications such as denoising or inpainting. For two dimensional images, local dictionaries take the form of small 2D patches, giving rise to the name patch-based dictionary learning where a collection of patch training examples can be taken to cover a full image or sets of images. Then the question arises as to the best method for reconstructing a new test image based on local dictionaries. As a first attempt, one may try to reconstruct every patch in an image independently. But then this may create artifacts between neighboring patches and does not give a global representation of the full image. To respect a global structure, a new methodology has become increasingly popular, known as convolutional sparse coding and dictionary learning, which unites a global image with local dictionaries through the convolution operator.

2.2.4 Convolutional Methods

The discrete two dimensional convolution between an $N \times N$ image I and a smaller $d \times d$ filter D is defined as:

$$(I * D)(x, y) = \sum_{m=-N/2}^{N/2} \sum_{n=-N/2}^{N/2} I(m, n) D(m - x, m - y). \tag{2.56}$$

In image processing this operation can be seen as applying a $d \times d$ filter centered at every pixel in an image, with the goal of smoothing or edge detection or feature extraction. For convolutional sparse coding, the correlation between the dictionary atoms at each image location are computed over the entire image. First, the convolutional sparse coding problem is:

$$\min_{\{x_k\}} \sum_{k=1}^{N_D} g(x_k) \text{ s.t. } \frac{1}{2} ||s - \sum_{k=1}^{N_D} D_k * x_k||_2^2 \leq \epsilon, \tag{2.57}$$

where each x_k is the size of the image s and is a sparse activation map which indicates the weighted locations of dictionary atom D_k. Then the convolutional dictionary learning problem is posed similarly but now with T training examples s_t and corresponding coefficients $x_{k,t}$ and a joint minimization over $\{x_{k,t}\}$ and D:

$$\min_{\{x_{k,t}\}, D} \sum_{t=1}^{T} \sum_{k=1}^{N_D} g(x_{k,t}) \text{ s.t. } \frac{1}{2} \sum_{t=1}^{T} ||s_t - \sum_{k=1}^{N_D} D_k * x_{k,t}||_2^2 \leq \epsilon, \ ||D_k||_2^2 \leq 1 \ \forall \ k, \tag{2.58}$$

There have been a number of recent algorithms to solve convolutional sparse cod-

ing/dictionary learning in an efficient manner. One popular strategy is exploiting the fact that a convolution in the Fourier domain reduces to the element-wise multiplication of the Fourier coefficients. The works of [74–76] explore various methods of efficiently exploiting the Fourier transform using ADMM but must take careful consideration of important boundary effects when going to and from the Fourier domain [77]. In a similar way, [78] convert the convolution into a large Toeplitz or circulant matrix and exploit the properties of its structure for efficient computation within ADMM. More recently, the work of [79–82] further analyze these circulant matrices and define new relationships between local sparsity in convolutional dictionaries and guarantees of global recovery. Alternatively, [83] uses FISTA and applies proximal gradient descent to the convolution directly.

We have thus outlined three machine learning topics, *sparse coding*, *compressed sensing*, and *dictionary learning*, that we will apply to dMRI data in the next three chapters and conclude with future work applying convolutional methods to sparse coding and compressed sensing in the final chapter. An over view of the ordered relationship between each of these topics is presented in Figure 2.21. While each has been applied to dMRI previously within a voxel-wise viewpoint, our major contribution in this thesis is to improve upon prior models within a global paradigm for dMRI processing and analysis.

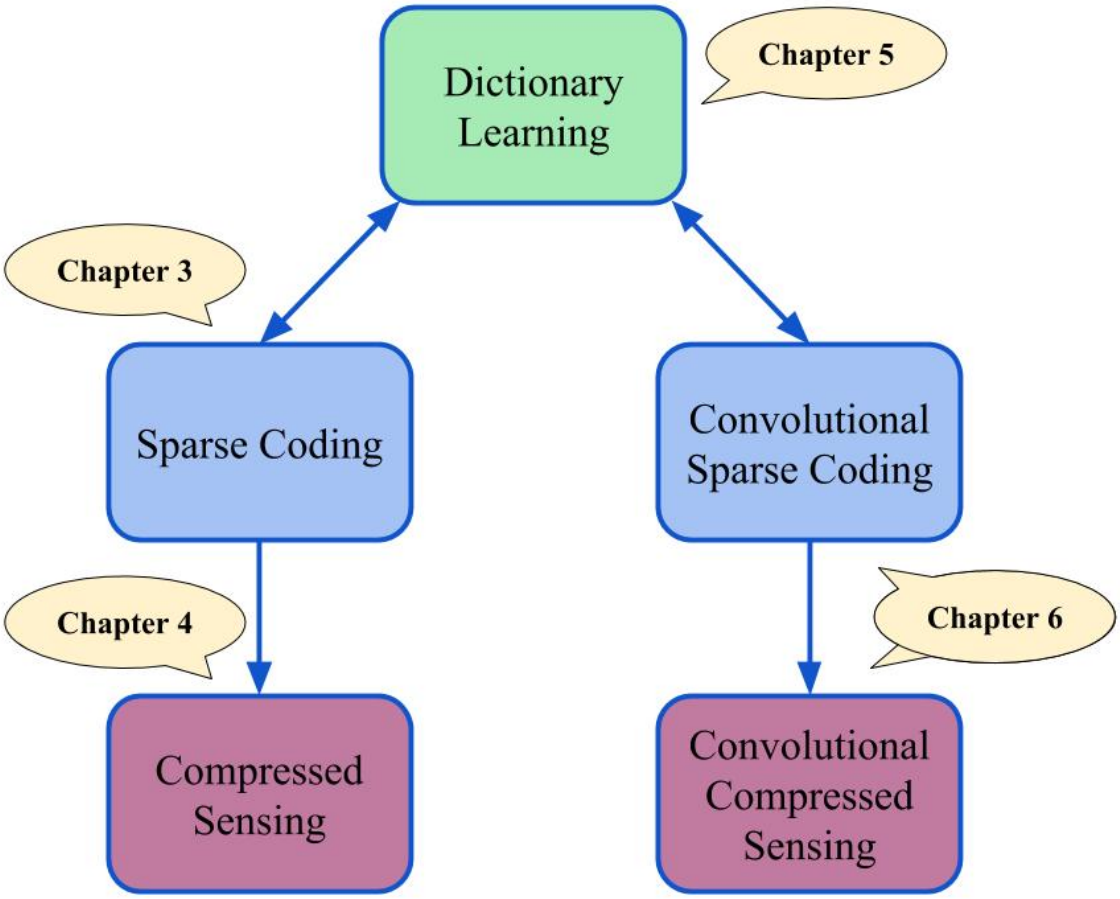


Figure 2.21: Overview of the relationships between each of our machine learning contributions.

Chapter 3

Spatial-Angular Sparse Coding

Finding sparse representations of data is important for a number of applications in signal processing (e.g., compressed sensing), image processing (e.g., dictionary learning), computer vision (e.g., face recognition), and biomedical imaging (e.g., MRI). In this chapter we show how sparse coding can be extended for the analysis of dMRI data. The contributions in this chapter have been published in [84, 85].

3.1 Introduction

As evidenced by our introduction in Chapter 2.1, dMRI data exhibit complex structure both in the spatial and angular domain. Nonetheless, much of the prior methods in sparse coding for dMRI rely on a voxel-wise viewpoint of the data and aim to find a sparse representation of the angular diffusion signals located at each

voxel in a brain volume. This makes the application of existing sparse coding methods to dMRI computationally efficient, as because a separate sparse coding problem must be solved at each voxel. However, because a non-zero signal must be modeled by at least one non-zero coefficient, voxel-wise or angular sparse coding will require at least one non-zero coefficient per voxel and thus for an entire volume the number of coefficients must be at least the number of total voxels. Herein lies a fundamental limitation of angular sparse coding. Some methods attempt to incorporate spatial regularization to enhance reconstructions of angular sparse coding based on sparse reconstruction work in MRI, but this will not improve the sparsity limit inherent in the problem formulation.

In this thesis, we propose a joint spatial-angular representation of dMRI that allows global sparsity levels to fall below one atom per voxel by exploiting redundancies in the spatial and angular domains, *jointly*. This can then open up the possibility to overcome the sparsity limits of the state of the art to achieve much higher acceleration rates for dMRI. A major challenge, however, is the computational complexity of solving a massive global sparse coding problem over large-scale dMRI data. Yet, by imposing that our global dictionary be separable over the spatial and angular domains we can greatly improve computational efficiency while preserving good sparsity levels for typical signals. One of our main contributions in this chapter is a set of novel adaptations of popular sparse coding algorithms to solve general large-scale sparse coding problems using separable dictionaries. Many of these algorithms will provide

insight into the related problems of compressed sensing and dictionary learning in the following chapters of this thesis. Our experiments on phantom and real HARDI brain data show that it is possible to achieve accurate global HARDI reconstructions with a sparse representation of less than one dictionary atom per voxel, exceeding the theoretical limit of the state of the art in sparse coding. Incorporating this *joint* spatial-angular sparse coding framework into a *joint* (k, q)-CS framework will be the subject of the work in Chapter 4.

The remainder of this chapter is organized as follows: In Section 3.2, we review state-of-the-art sparse coding and CS dMRI methods and illustrate the limitations of their performance on a phantom HARDI dataset. In Section 3.3, we present our joint spatial-angular dMRI representation and formalize the global spatial-angular sparse coding problem. Then, in Section 3.4, we develop and compare a set of novel sparse coding algorithms using separable dictionaries to efficiently solve our large-scale global optimization. Finally, in Section 3.5 we provide experimental results showing the performance of our method over the state-of-the-art and conclude with a discussion in Section 3.6.

3.2 State of the Art in Sparse Coding

Existing sparse coding methods for dMRI can be divided in three main categories: spatial, angular, and a combination of spatial and angular. While spatial

		Sparse Coding			
		Spatial	Angular	Spatial + Angular	Joint Spatial-Angular
CS	k	[60]			
	q		[61, 86–93]		
	(k, q)			[62, 94–97]	Proposed

Table 3.1: Summary of the state-of-the-art dMRI sparse reconstruction methods organized by domains of sparse coding and CS subsampling. The literature has provided a natural extension from k-CS in MRI using spatial sparse coding to q-CS in dMRI angular sparse coding. However, for (k, q)-CS, the state of the art enforce sparsity in the spatial and angular domains separately, (called "Spatial + Angular" Sparse Coding). In contrast, the proposed work considers a joint spatial-angular representation for sparse coding which is a more natural model for joint (k, q)-CS.

sparse coding (MRI) and angular sparse coding (dMRI) can be written in the classical framework described in Chapter 2.2.1, and classically extended to compressed sensing in k-space and q-space respectively, the combination of spatial and angular sparse coding becomes more involved. The next subsections summarize state of the art methods for angular sparse coding and Table 3.1 organizes the recent literature into their respective categories of spatial/angular sparse coding and their domains for compressed sensing.

3.2.1 Angular (Voxel-Wise) Reconstruction

A dMRI can be modeled as a 6D signal $\mathcal{S}(v, q)$, where $v \in \Omega \subset \mathbb{R}^3$ is the location of a voxel in the 3D spatial domain Ω and $q \in \mathbb{R}^3$ is a point in the so-called q-space.[1] A dMRI signal is measured at a discrete number of voxels, V, and a discrete number of q-space points, G. While dMRI signals can be viewed as a set of G

[1]The q-space is the frequency domain associated with the angular domain, while the k-space is the frequency domain associated with the spatial domain.

diffusion weighted images (DWIs) or volumes, the most common view-point for dMRI processing and analysis is voxel-wise, i.e. for each voxel $v \in \Omega$, we acquire a vector of G diffusion measurements $\mathcal{S}(v, q_g)_{g=1}^{G} = s_v(q_g)_{g=1}^{G} = s_v$ at points q_g in 3D q-space. This interpretation is most common for modeling because a major goal of dMRI reconstruction is to estimate 3D probability distribution functions (PDFs) of fiber tract orientation at each voxel. Accordingly, the signal vector s_v is represented by a q-space, or angular, dictionary $\Gamma = [\Gamma_i(q)]_{i=1}^{N_\Gamma}$, with N_Γ atoms, such that

$$s_v = \Gamma a_v. \tag{3.1}$$

where a_v are the angular coefficients for voxel v. The dMRI literature has produced a wide array of dMRI reconstruction algorithms for different acquisition protocols, an artillery of q-space bases and varying models for estimating PDFs of fiber orientation. As discussed in Section 2.1.4, the vast majority of research reconstructs q-space signals in each voxel with a q-space basis [98] while adding a set of constraints $\mathcal{C}$ on the coefficients to enforce desirable properties such as non-negativity of PDFs [12, 54] or angular smoothing [99]. More specifically the angular coefficients in (3.1) are found by solving:

$$a_v^* = \arg\min_{a_v} \frac{1}{2}||\Gamma a_v - s_v||_2^2 \quad \text{s.t.} \quad a_v \in \mathcal{C}. \tag{3.2}$$

The constraint of particular interest in this thesis is that of enforcing sparsity on the coefficients of the reconstruction, known as *Sparse Coding*, which has applications in

CS as well as super-resolution [100] and de-noising [101].

3.2.2 Angular (Voxel-Wise) Sparse Coding

Sparse coding is a reconstruction problem which seeks a sparse representation, i.e. a coefficient vector with few nonzero elements. Given a sparsifying q-space basis Γ for which the dMRI signal in each voxel is expected to have a sparse representation, the angular (voxel-wise) sparse coding problem can be formulated as:

$$a_v^* = \arg\min_{a_v} \frac{1}{2}||\Gamma a_v - s_v||_2^2 \quad \text{s.t.} \quad ||a_v||_0 \leq K_v, \tag{3.3}$$

where $||a_v||_0$ counts the number of nonzero elements of vector a_v, and K_v is the sparsity level at voxel v. This problem is known to be NP-hard, and therefore the two main methodologies to tackle (3.3) are to a) approximate a solution using greedy algorithms such as Orthogonal Matching Pursuit (OMP) [67] or b) replace the L_0 semi-norm by its convex relaxation, the L_1 norm, and solve either the Basis Pursuit or LASSO problem given by:

$$a_v^* = \arg\min_{a_v} \frac{1}{2}||\Gamma a_v - s_v||_2^2 + \lambda||a_v||_1 \tag{3.4}$$

using algorithms such as Alternating Direction Method of Multipliers (ADMM) [102] or Fast Iterative Thresholding Algorithm (FISTA) [103], where λ is the trade-off

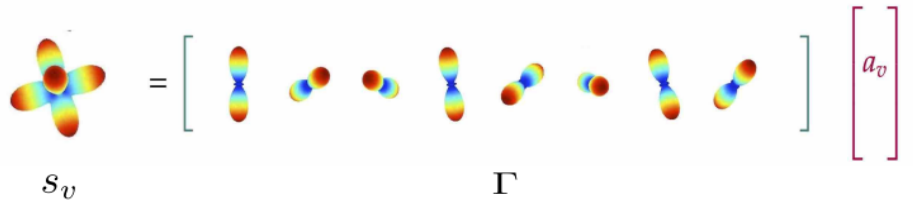


Figure 3.1: Illustration of voxel-wise angular HARDI representation a_v using a sparsifying dictionary Γ.

parameter between data fidelity and sparsity. Angular sparse coding and q-CS have been widely researched for dMRI to reduce long acquisition times. Many groups have done extensive work choosing sparsifying q-space bases [61, 90, 104], developing dictionary learning methods [88, 105–111], and testing q-space subsampling schemes for DSI [86, 108, 112–114], MS-HARDI [89, 110, 115–117], HARDI [15, 44, 91, 118, 119] and DTI [120] with promising results in sparsity and measurement reduction for clinical tractography [121,122]. However, a major limitation for this family of methods is that the sparsest possible representation of an entire dMRI dataset can be no less than the number of voxels since $||a_v||_0 \geq 1 \; \forall \; v \in \Omega$. In CS applications, this induces fundamental limitations in the amount of subsampling factors that may be achievable in q-space. In practice, the global sparsity level will be much greater than the number of voxels, to account for noise. For example the work of [14, 15] report an average sparsity level of 6 to 10 atoms per voxel. The methods presented in the next section attempt to improve upon these results by exploiting spatial redundancies and reducing measurements in k-space.

3.2.3 Angular Sparse Coding with Spatial Regularization

Incorporating spatial information into voxel-wise reconstruction is a well utilized technique for increasing the accuracy of reconstruction. The following is a general formulation for including spatial regularization into the angular sparse coding problem:

$$A^* = \arg\min_A ||\Gamma A - S||_F^2 + \lambda||A||_1 + \mathcal{R}(A), \tag{3.5}$$

where $S = [s_1 \dots s_V] \in \mathbb{R}^{G\times V}$ is the concatenation of signals $s_v \in \mathbb{R}^G$ sampled at G gradient directions over V voxels, $A = [a_1 \dots a_V] \in \mathbb{R}^{N_\Gamma \times V}$ is the concatenation of angular coefficients and $\mathcal{R}(A)$ is a spatial regularizer that depends on the angular representation A. Here $||X||_F = \sqrt{\sum_i \sum_j |X_{i,j}|^2}$ is the Frobenius norm and $||X||_1 = \sum_i \sum_j |X_{i,j}|$ is the 1-norm taken over all elements of the matrix. In particular when $\mathcal{R} = 0$, this reduces to solving (3.4) for each voxel. When $\lambda = 0$ and $\mathcal{R}(A) = \sum_i \sum_{j\in\mathcal{N}_i} \|a_i - a_j\|^2$ (Laplacian regularization), where $\mathcal{N}_i$ is a local spatial neighborhood of voxel i, this is the general non-sparse reconstruction with spatial coherence (see [32] for example). Some have found incorporating both angular sparsity $\lambda||A||_1$ and spatial coherence $\mathcal{R}(A)$ beneficial for applications such as de-noising [99, 101, 123, 124] and tractography [125].

Spatial regularization within sparse coding is more prominently used for the application of reducing redundancies for CS. For example, [45, 87, 126] enforce spatial

smoothing for q-CS while [92, 93] combine q-CS, with super-resolution reconstruction of the spatial domain. To further accelerate dMRI, the recent work of [127] combines CS with parallel imaging but reconstructs the signals in first in k-space and then in q-space, separately. A joint (k, q)-space reconstruction is important for maintaining coherence throughout the dataset. As such, the works of [62, 94–97, 128] combine k- and q-CS by adding a data fidelity term for k-space subsampling and an additional spatial sparsity term. In total, however, while each of these works may be applied to different diffusion models and acquisition protocols testing various subsampling schemes, sparsifying transforms and dictionaries, each are based on an angular representation of dMRI data, A. In fact, they stem from the same optimization problem formulation (3.5) with

$$\mathcal{R}(A) = \xi ||\Psi(\Gamma A)||_1, \tag{3.6}$$

where $\xi \geq 0$ is an additional trade-off weighting parameter, and $\Psi(\cdot)$ is a sparsifying transform (or dictionary) applied to the spatial domain such as wavelets or the finite difference gradient operator, leading to the usual total variation (TV) norm. In (3.6), ΓA is a reconstruction of the signal S based on the angular representation A.

While adding these spatial and angular sparsity terms may exploit redundancies in both the spatial and angular domains, because they are separate disjoint terms the minimal global sparsity level will be still limited by the size of the data since $||A||_0$ should be greater than V and $||\Psi(\Gamma A)||_0$ should be greater than G. Indeed, when $||A||_0 < V$, there must exist voxels v such that $a_v = 0$, leading to a zero valued signal

s_v (column of S) in that voxel. Likewise, when $||\Psi(\Gamma A)||_0 < G$, there must exist some gradient directions, q_g, such that the signal in the entire volume s_q (rows of S) equals zero. This becomes a problem because zero valued signals are not physically representative of real dMRI data. This also becomes a heuristic limitation of prior methods for appropriately choosing trade-off parameters λ and ξ that result in a physically accurate sparsity level. In the next section we will explicitly show this sparsity limitation on phantom data.

Table 3.1 organizes the recent literature's usages of sparse coding and CS for dMRI and places the proposed work in context compared to the state of the art. There we use the term "Spatial + Angular" Sparse Coding to emphasize that the state of the art perform both spatial and angular sparse coding, but not jointly. In light of a joint (k, q)-CS, a disjoint set of spatial and angular sparsifying transforms may also not be the most natural choice from a classical CS perspective.

3.2.4 Limitations of Angular Representations for Sparse Coding

We illustrate the limitations of sparse coding using a per-voxel angular representation on a HARDI phantom dataset with $V = 50 \times 50$ and $G = 64$ gradient directions (the same data is used in our experiments in Section 4.6). First, we solve (3.5) with $\mathcal{R}(A) = 0$, showing qualitative reconstruction results in Figure 3.2, for various spar-

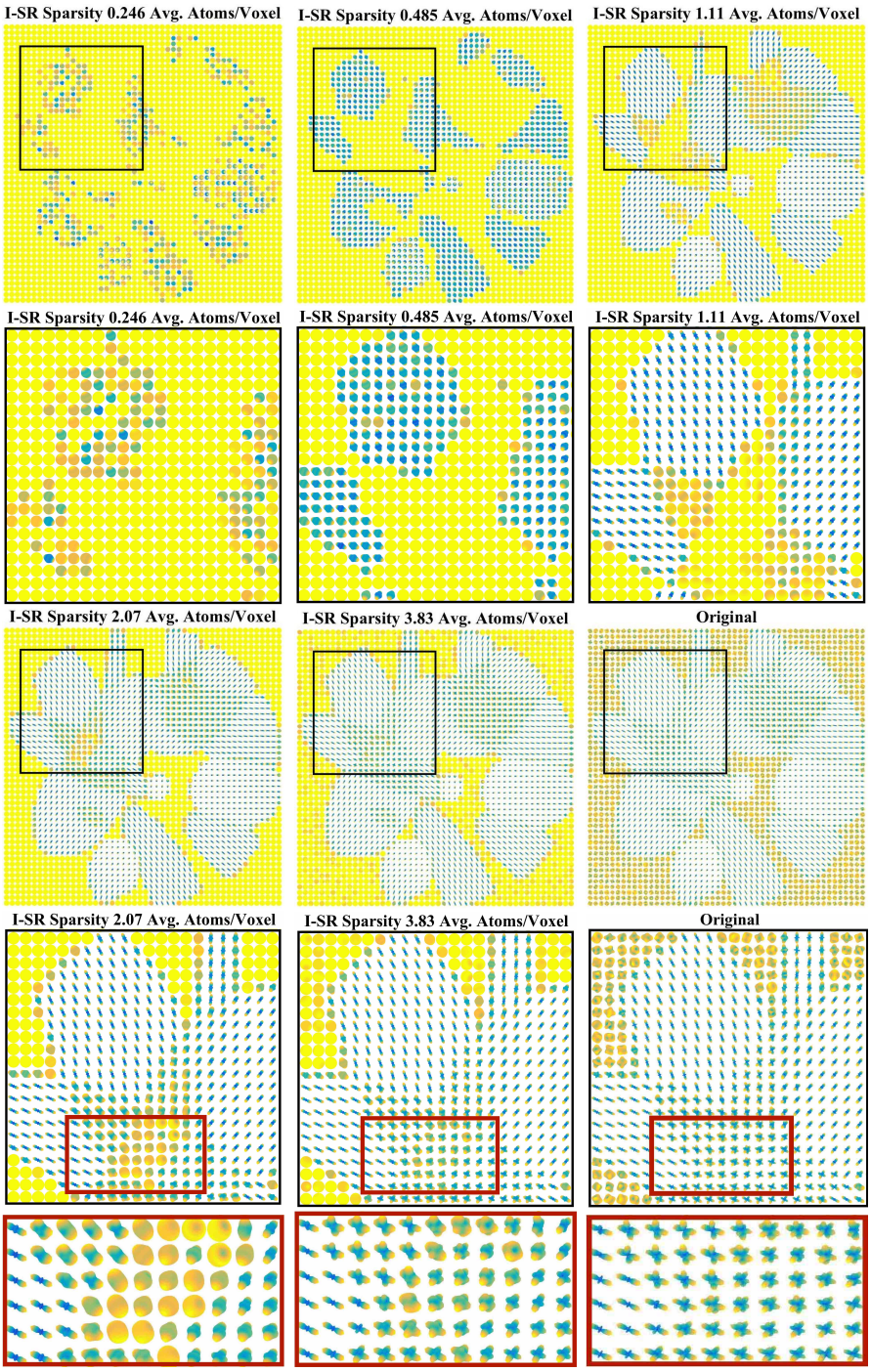


Figure 3.2: Qualitative demonstration of state-of-the-art sparse coding limitations (3.5) with the spherical ridgelets (SR) dictionary for 5 different spatial-angular sparsity levels compared to the original signal (bottom right) with ROI closeups underneath. For high spatial-angular sparsity levels (top left, middle), voxels with complex signals are forced to zero (yellow spheres). Regions with crossing fibers are unable to be accurately reconstructed even when using an average of 2.07 atoms/voxel. The label I-SR refers to Identity-SR, explained in the next section.

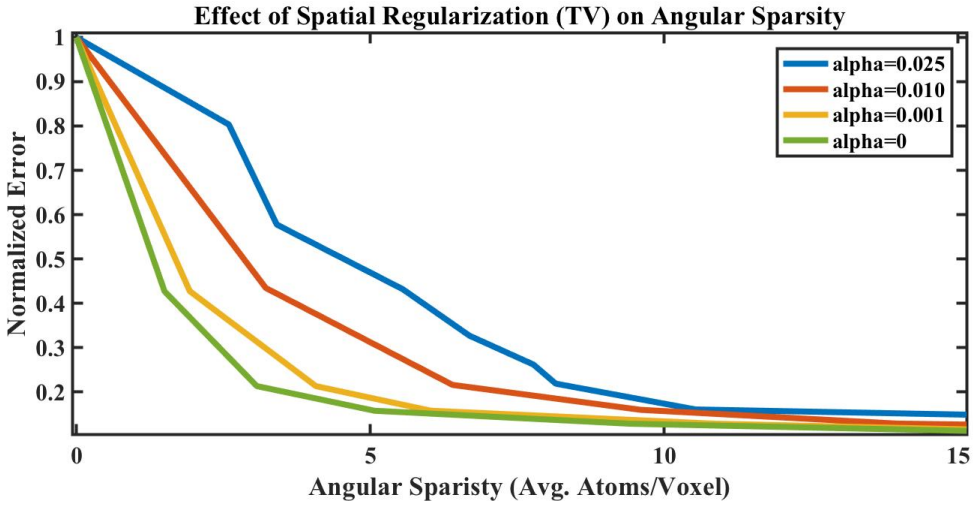


Figure 3.3: Reconstruction error vs. the average number of angular dictionary atoms per voxel using spatial regularization for the HARDI phantom data. As α, the relative weight of spatial regularization (TV) in (3.6), increases, the average number of angular atoms increases for a given reconstruction error. This suggests that sparser solutions for angular sparse coding can be achieved without spatial regularization, although using adequate spatial regularizers can improve the qualitative aspect of the reconstructed signal, in particular for noisy inputs.

sity levels given by the value of λ. Our second result considers the effect of spatial regularization $\mathcal{R}(A) \neq 0$ on the amount of angular sparsity as a function of the reconstruction error in Figure 3.3.

For this setting, we choose angular basis Γ to be the well performing overcomplete spherical ridglet (SR) dictionary [14, 44, 118]. Figure 3.2 shows the ODF estimations (computed using the spherical wavelets [44]) from the sparse signal reconstruction for various sparsity levels compared to the ODFs estimated from the original signal, as well as close-ups of a region of interest (ROI) containing ODFs with complex crossings of 2, 3 and 4 fibers. In order to compare spatial-angular sparsity levels we are interested in the average number of active dictionary atoms over all voxels,

i.e. $||A||_0/V$. We use 5 different values of λ which gives us average spatial-angular sparsity levels of $0.246, 0.485, 1.11, 2.07$, and 3.84 atoms per voxel. As expected, when $||A||_0/V < 1$ (see top left/middle), many voxels are forced to zero (as indicated by yellow spheres in Figure 3.2). This is especially true for isotropic signals surrounding the fiber tracts. Also as expected, when $||A||_0/V \approx 1$, (see top right) many of the complex signals in the fiber crossing ROI are pushed to zero. This model requires an average of close to $||A||_0/V = 4$ atoms per voxel to achieve nearly accurate signal reconstruction (bottom middle). In fact, the actual number of coefficients per voxel to accurately represent typical dMRI data with angular bases is substantially higher. We illustrate this in Figure 3.4 which shows the number of atoms used to represent the HARDI signals in each voxel for the reconstructions in Figure 3.2. The bottom right image shows the ground truth number of fibers crossing in each voxel. This experiment demonstrates that voxels containing crossing fibers are forced to zero atoms when the average number of atoms per voxel is very small and tend to 6-12 atoms for accurate reconstruction when the sparsity level is decreased. This is consistent with the reports of [14, 15] for the SR dictionary.

Next, we explore the effect of adding spatial regularization $\mathcal{R}$ to the angular sparsity penalty, as a function of the reconstruction error. As a common spatial regularizer used in the literature, we consider for $\mathcal{T}$ in (3.6) the finite difference (gradient) operator $\mathcal{T} = \nabla := [\partial_x, \partial_y, \partial_z]$ and the corresponding isotropic TV norm given by $||\nabla(X)||_{2,1} = ||\sqrt{|\partial_x X|^2 + |\partial_y X|^2 + |\partial_z X|^2}||_1$. This leads to the new optimization

problem

$$A^* = \arg\min_A ||\Gamma A - S||_F^2 + \lambda||A||_1 + \alpha\lambda||\nabla(\Gamma A)||_{2,1}, \tag{3.7}$$

for various λ and $\alpha \geq 0$, the relative weight of spatial regularization. Figure 3.3 shows the effect of nonzero α on angular sparsity compared to the case of $\alpha = 0$ ($\mathcal{R} = 0$) on a small 30×30 segment of the phantom HARDI data. As we can see, in all cases, the minimal sparsity for accurate reconstruction does not go below the limit of 5 atoms per voxel. In addition, increasing the relative weight of the TV norm spatial regularization actually results in an increase in angular sparsity for a given reconstruction error. In a sense, this is not surprising since the additional regularizer $\mathcal{R}$ will enforce spatial smoothness of the reconstructed signal (which can be beneficial for noisy data and in compressed sensing scenarios) but cannot improve the resulting sparsity of the solution which is still represented by a set of coefficients per voxel in the angular basis Γ. As the goal of this paper is sparse coding, i.e finding sparsest possible representations of full HARDI data, in our later experimental comparisons, we will be using $\mathcal{R} = 0$ when referring to state-of-the-art reconstruction.

In the following section, we present our global spatial-angular representation of dMRI which allows for the possibility to achieve accurate reconstruction with sparsity levels below the number of voxels, unachievable with an angular representation alone.

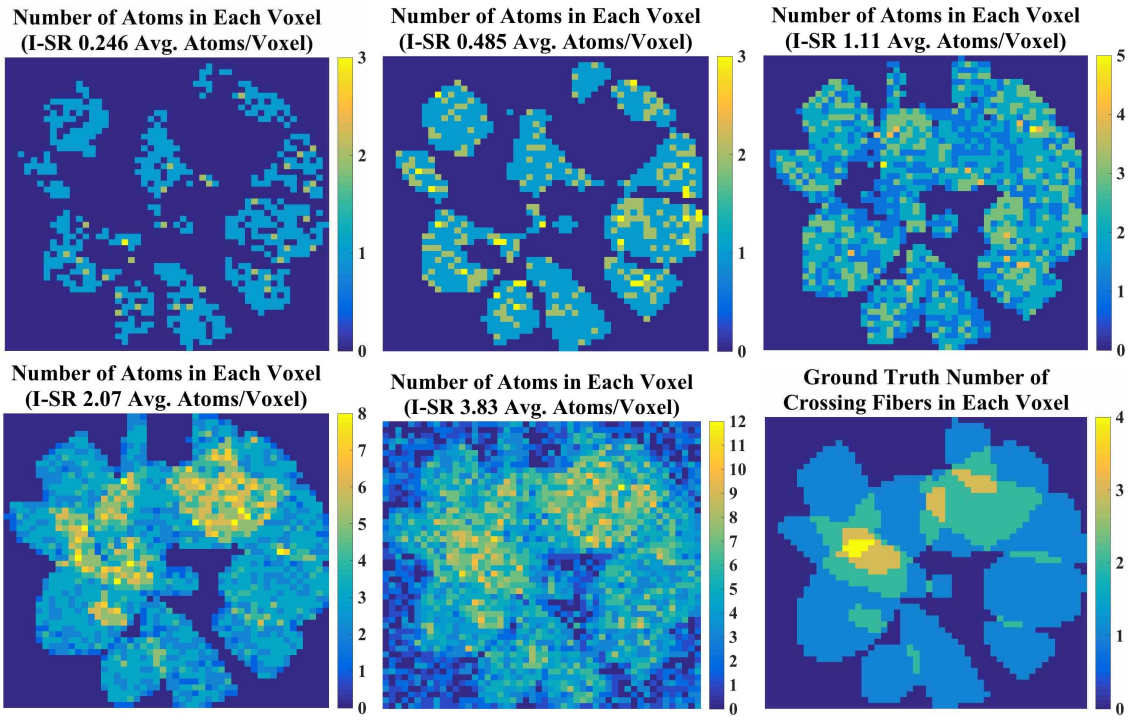


Figure 3.4: Number of atoms found in each voxel corresponding to the 5 levels of spatial-angular sparsity in Figure (3.2). The bottom right figure shows the ground truth number of fibers crossing in each voxel to illustrate the complexity of each angular signal in relation to how many atoms are needed to sparsely model them. Crossing fiber signals are either forced to zero for high spatial-angular sparsity levels (see: top row) or require between 3-5 atoms for single fiber signals (see: avg. sparsity 1.11 and 2.07) and 6-12 for double and triple crossing fiber signals (see: avg. sparsity 3.83). The label I-SR refers to Identity-SR, explained in the experiments Section 3.5.

3.3 Joint Spatial-Angular dMRI Representation

To overcome the sparsity limits of an angular representation, we propose to model a dMRI signal $\mathcal{S} : \Omega \times \mathbb{R}^3 \to \mathbb{R}$ globally with a joint spatial-angular dictionary, say

$\varphi(v,q)$, such that

$$\mathcal{S}(v,q) = \sum_k c_k \varphi_k(v,q) \tag{3.8}$$

with a single set of global coefficients $c = [c_k]$. A global dictionary allows us to find global representations with sparsity levels below the number of voxels without forcing some voxels to have zero signal. In fact, the sparsest possible representation would be the absolute limit of 1 nonzero coefficient c_k, and so we find ourselves in a unrestricted setting for global sparse coding. To set up the spatial-angular sparse coding problem, we let $s \in \mathbb{R}^{GV}$ be the vectorization of $\mathcal{S}(v,q)$ where for $v = 1 \dots V$ we stack the q-space signals, $s_v \in \mathbb{R}^G$, and $\Phi_k \in \mathbb{R}^{GV}$ be the vectorization $\varphi_k(v,q)$ to build the global dictionary $\Phi = [\Phi_1 \dots \Phi_{N_\Phi}] \in \mathbb{R}^{GV \times N_\Phi}$, with N_Φ atoms. Then, to find a globally sparse c, we can solve the L_0 minimization problem:

$$c^* = \arg\min_c \frac{1}{2}||\Phi c - s||_2^2 \;\text{ s.t. }\; ||c||_0 \leq K, \tag{$P0vec$}$$

for a sparsity level K or the LASSO problem:

$$c^* = \arg\min_c \frac{1}{2}||\Phi c - s||_2^2 + \lambda||c||_1, \tag{$P1vec$}$$

where $\lambda > 0$ is the sparsity trade-off parameter. However, typical dMRI contains on the order of $V \approx 100^3$ voxels each with $G \approx 100$ q-space measurements for a total of $100^4 = 100$ million signal measurements ($|s| \approx 10^8$). Since many sparse

coding applications often utilize dictionaries that are over-redundant, this leads to a massive matrix Φ with 100^4 rows and over 100^4 columns ($|\Phi| \approx 10^{16}$). For some datasets, even committing Φ to memory is prohibitive. Therefore solving this large-scale global dMRI sparse coding problem using traditional solvers like OMP to approximate ($P0vec$) or ADMM and FISTA to solve ($P1vec$), prove intractable.

To address this challenge, we introduce additional structure on the dictionary atoms by considering separable functions over Ω and $\mathbb{R}^3$, namely a set of atoms of the form $\{\varphi_k(v,q)\} = \{\psi_j(v) \otimes \gamma_i(q)\}$, where $\{\psi_j(v)\}$ is a spatial basis for the space of functions from $\Omega \to \mathbb{R}$ and $\{\gamma_i(q)\}$ is an angular basis for the space of functions from $\mathbb{R}^3 \to \mathbb{R}$ and $\otimes$ is the Kronecker product. In discretized form for V voxels and G gradient directions, with $\Psi \in \mathbb{R}^{V \times N_\Psi}$ and $\Gamma \in \mathbb{R}^{G \times N_\Gamma}$, the matrix $\Phi = \Psi \otimes \Gamma$ is of the form:

$$s = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_V \end{pmatrix} = \begin{pmatrix} \Psi_{1,1}\Gamma & \Psi_{1,2}\Gamma & \cdots & \Psi_{1,N_\Psi}\Gamma \\ \Psi_{2,1}\Gamma & \Psi_{2,2}\Gamma & \cdots & \Psi_{2,N_\Psi}\Gamma \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{V,1}\Gamma & \Psi_{V,2}\Gamma & \cdots & \Psi_{V,N_\Psi}\Gamma \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N_\Psi N_\Gamma} \end{pmatrix} = \Phi c. \tag{3.9}$$

Figure 3.5 illustrates the Kronecker structure of spatial-angular atom Φ_k. We can see that by representing a dMRI signal with this type of global spatial-angular atom, one can model an entire region of the brain with as few as a single atom instead of angular atoms at every voxel.

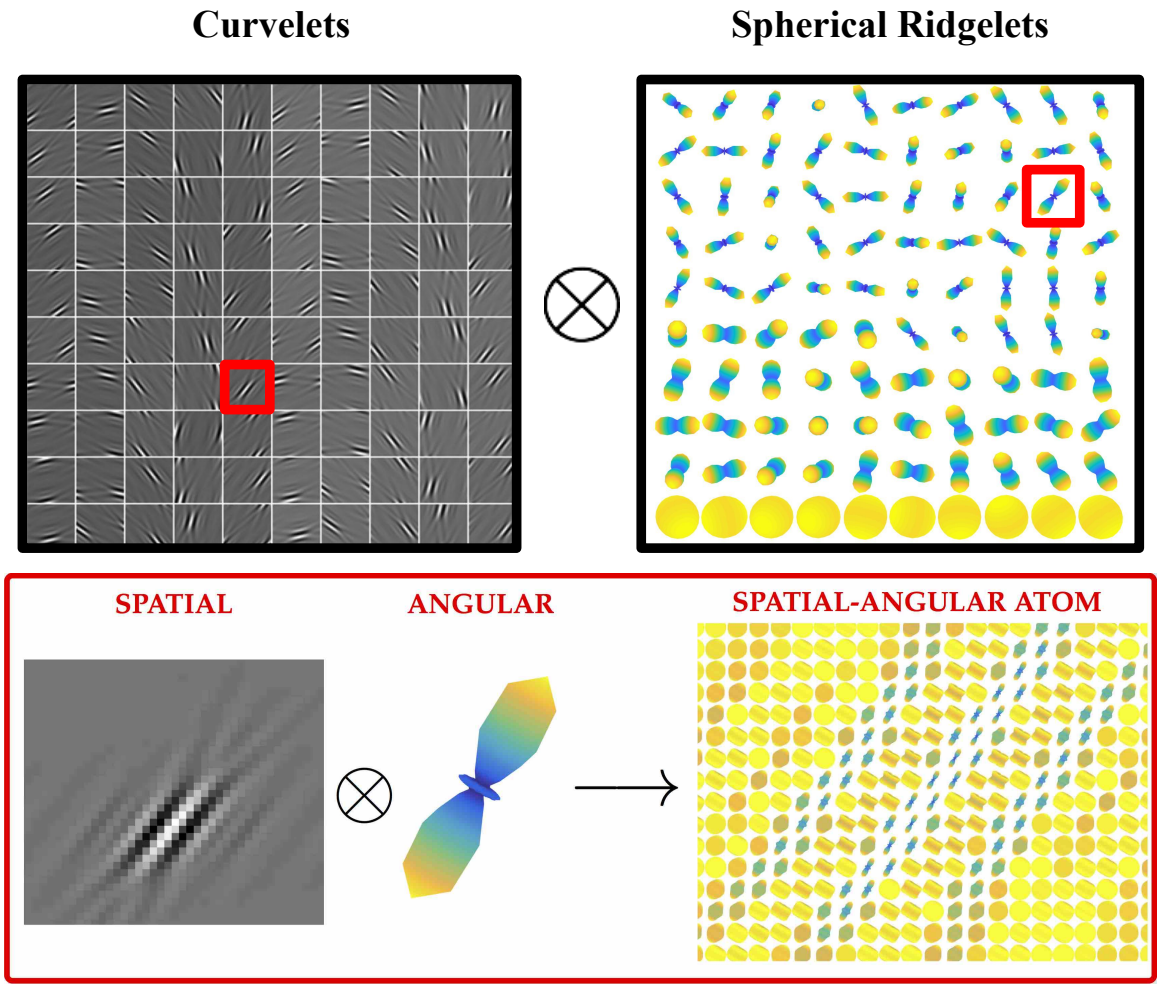


Figure 3.5: Top: A separable spatial-angular dictionary composed of the Kronecker product between curvelets Ψ and spherical ridgelets Γ. A pair of spatial and angular atoms are highlighted in red and zoomed in below. Bottom: An example construction of a single spatial-angular basis atom Φ_k (right) by taking the Kronecker product of Ψ_j (left) and Γ_i (middle), i.e. $\Psi_j \otimes \Gamma_i = \Phi_k$. With this particular combination of spatial (curvelet Candes:MMS06) and angular (spherical wavelet TristanVega:MICCAI11) atoms, we can see that it may be possible to represent an entire fiber tract with very few spatial-angular atoms.

A motivating model for this separable structure for dMRI is as follows: first, as is traditionally done, the signal at each voxel $v \in \Omega$ is written as a linear combination

of angular basis functions $\{\Gamma_i(q)\}$:

$$\mathcal{S}(v,q) = \sum_{i=1}^{N_\Gamma} a_i(v)\Gamma_i(q). \tag{3.10}$$

Then, we notice that each spherical coefficient $a_i(v)$ forms a 3D volume and so can be written as a linear combination of spatial basis functions $\{\Psi_j(v)\}$:

$$a_i(v) = \sum_{j=1}^{N_\Psi} c_{i,j}\Psi_j(v). \tag{3.11}$$

Combining (3.10) and (3.11) we arrive at our proposed separable spatial-angular dictionary

$$\mathcal{S}(v,q) = \sum_{i=1}^{N_\Gamma}\sum_{j=1}^{N_\Psi} c_{i,j}\Psi_j(v)\Gamma_i(q), \tag{3.12}$$

When stacking each s_v in a large vector, (3.12) results in the Kronecker product in (3.9), $s = (\Psi \otimes \Gamma)c$. Alternatively, when writing $S = [s_1, \ldots, s_V]$ as a matrix, (3.12) results in the equivalent matrix form:

$$S = \Gamma C \Psi^\top. \tag{3.13}$$

Table 3.2 summaries the dimensions of the vector and matrix variables and Figure 3.6 illustrates the Kronecker decompositions in the vector and matrix forms.

Decomposing signals into Kronecker (or more general multi-tensor) structures has been well researched to increase algorithmic efficiency by reducing computations

	Signal		Coefficients		Dictionaries		
Variable	s	S	c	C	Φ	Γ	Ψ
Dimensions	GV	$G \times V$	$N_\Gamma N_\Psi$	$N_\Gamma \times N_\Psi$	$GV \times N_\Gamma N_\Psi$	$G \times N_\Gamma$	$V \times N_\Psi$

Table 3.2: Sparse coding variable dimensions, where G (≈ 100) is the number of gradient directions in q-space, V ($\approx 100^3$) is the number of voxels in the volume, N_Γ ($\gtrsim 100$) is the number of atoms of the angular dictionary Γ, and N_Ψ ($\gtrsim 100^3$) is the number of atoms of the spatial dictionary Ψ.

to the smaller, separate domains. Many research groups have exploited properties of the Kronecker product, when solving problem types of the form of ($P0vec$) and ($P1vec$) for computational efficiency of larger sparse coding [129], dictionary learning [130] and CS [131] applications. The work of [132] has applied multi-tensor sparse coding methods on dMRI data for the application of fiber tract data compression. In particular, a Kronecker Orthogonal Matching Pursuit (Kron-OMP) [133] has been utilized to solve ($P0vec$). Although Kron-OMP becomes much more efficient than the classical OMP [67], the problem is not entirely separated into smaller domains, and the computationally expensive Φ matrix is still built explicitly. For large-scale problems like that of dMRI reconstruction, solving ($P0vec$) or ($P1vec$) even with a Kronecker structure dictionary remains largely intractable/expensive for memory and computation time.

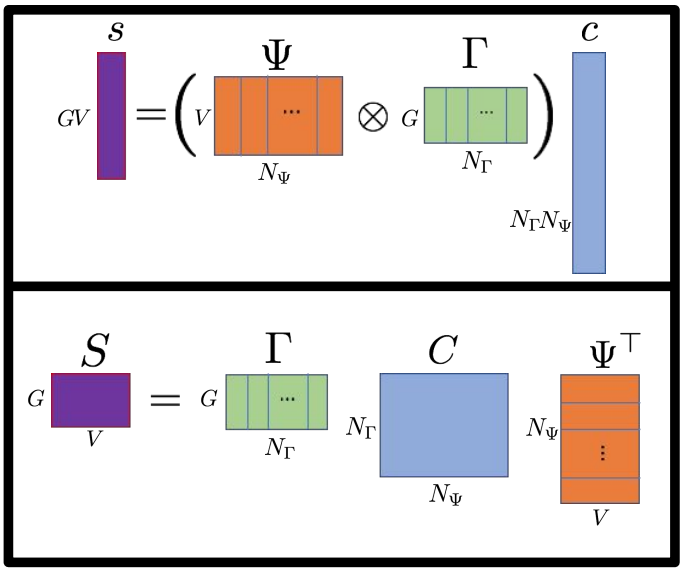


Figure 3.6: Equivalent vector form (top) and matrix form (bottom) for the Kronecker decomposition of a signal. We propose to use the matrix form which provides a more compact representation for signals of large size and exploits the full separability of the Kronecker product, reducing matrix multiplication complexity from $O(GVN_\Gamma N_\Psi)$ to $O(GVN_\Gamma)$.

In this thesis, we propose to use the matrix form (3.13) which allows us to avoid the expensive uses of Φ and fully reduce computational complexity to the smaller separable basis domains of Γ and Ψ (consult Table **??** for a review of variable dimensions). In particular we develop efficient algorithms to solve the completely separable spatial-angular sparse coding problems:

$$C^* = \arg\min_C \frac{1}{2}||\Gamma C\Psi^\top - S||_F^2 \quad \text{s.t.} \quad ||C||_0 \leq K \qquad (P0mat)$$

and

$$C^* = \arg\min_C \frac{1}{2}||\Gamma C\Psi^\top - S||_F^2 + \lambda||C||_1. \qquad (P1mat)$$

This becomes a general optimization to solve large-scale sparse coding problems with separable dictionaries and can also be extended to the tensor setting.

As an important note, this matrix formulation is a generalization of the voxel-wise angular sparse coding problem (3.5) in the special case of $\Psi = \mathrm{I}_V$, the $V \times V$ identity matrix, with $C \equiv A$. We use the identity as a choice for Ψ in the experiments of Section 3.5 when comparing the performance of purely angular sparse coding with our proposed framework[2].

3.4 Efficient Kronecker Sparse Coding Algorithms

In what follows we present three novel adaptations of existing sparse coding algorithms for solving large-scale sparse coding problems with a Kronecker dictionary structure. These are Kronecker extensions of OMP (Section 3.4.2), ADMM/Dual ADMM (Section 3.4.3 and Section 3.4.4), and FISTA (Section 3.4.5). We compare these to an existing Kronecker OMP algorithm proposed in [133] (Section 3.4.1). We compare these algorithms in terms of complexity for various types of bases in Section 3.4.6 and show experimental time comparisons in Section 3.5. (See Chapter 2.2.1.2 for a review of the classical OMP, FISTA and ADMM algorithms.)

[2]Using $\Psi = \mathrm{I}_V$ identity with spherical ridgelets (SR) we adopt the notation I-SR for the dictionary used in the state-of-the-art illustration Figure 3.2 and Section 3.5.

3.4.1 Kronecker OMP

To approximate a solution to the L_0 problem ($P0vec$), Orthogonal Matching Pursuit (OMP) [67] is a popular greedy algorithm that iteratively selects the atom that is most correlated with the signal, orthogonalizes it to the previously selected atoms by solving a least squares optimization, and selects the next atom that is most correlated with the resulting residual. For the case of a Kronecker structured basis, a Kronecker OMP (Kron-OMP) algorithm has been previously proposed [129, 133] that reduces computations of solving the least squares subproblem in each iteration by exploiting properties of the Kronecker product. This form of Kron-OMP, however, represents the signal, coefficients, and basis atoms in vector form providing a solution to ($P0vec$). In Algorithm 4 we rewrite the Kron-OMP algorithm adapted to the structure of our problem, where $vec(\cdot)$ and $mat(\cdot)$ convert matrices to vectors and vice versa. The main complexity gain of Kron-OMP over the vector OMP is obtained by separating the effects of Γ and Ψ when computing the maximally correlated atoms with the residual, $|\Gamma^\top R\Psi|$ (See Algorithm 4 Step 1) with complexity $O(N_\Gamma GV + GN_\Gamma N_\Psi)$ instead of computing $|\Phi^\top r|$ with complexity $O(N_\Gamma N_\Psi GV)$. The other gain is in solving the least squares problem (See Algorithm 4 Step 3) by exploiting properties of the Kronecker product ($A \odot B = [a_1 \otimes b_1, \ldots, a_N \otimes B_N]$) to compute a rank-1 update. However, the only real improvement on complexity is in memory since Φ can be built atom by atom from the columns of Γ and Ψ instead of storing the entire matrix. The rank-1 update remains $O(k^2)$ for both vector and Kron-OMP. In the next section

we present an alternative Kron-OMP algorithm that reduces complexity further by exploiting the full separability of the dictionary.

Algorithm 4 Kron-OMP

Choose: K, ϵ.
Initialize: $k = 1,\ \mathcal{I}^0 = \emptyset,\ \mathcal{J}^0 = \emptyset,\ R_0 = S,\ s = vec(S)$.
while $k \leq K$ and error $> \epsilon$ **do**
1: $[i^k, j^k] = \arg\max_{[i,j]} |\Gamma_i^\top R_k \Psi_j|$;
2: $\mathcal{I}^k = [\mathcal{I}^{k-1},\ i^k]; \mathcal{J}^k = [\mathcal{J}^{k-1},\ j^k]; \mathcal{A}^k = (\mathcal{I}^k, \mathcal{J}^k)$;
3: $c_k = \arg\min_c \frac{1}{2}||(\Gamma_{\mathcal{I}^k} \odot \Psi_{\mathcal{J}^k})c - s||_2^2$;
4: $R_k = mat(s - (\Gamma_{\mathcal{I}^k} \odot \Psi_{\mathcal{J}^k})c_k)$;
5: $k \leftarrow k + 1$;
end while
Return: $\mathcal{A}^K, c_K$

3.4.2 Kronecker OMP with Projected Gradient Descent

In what follows, we develop a novel form of Kronecker OMP which solves the separable ($P0mat$) instead of ($P0vec$). This allows us to reduce computation by not building columns of Φ and not repeating individual atoms of Γ or Ψ. Instead, indices of Γ and Ψ are updated only when they each have not been chosen before, fully exploiting the separability of the dictionary. Given the previous sets of respective of indices $\mathcal{I}^{k-1}$ and $\mathcal{J}^{k-1}$, we update sets by following $\mathcal{I}^k = [\mathcal{I}^{k-1}\ i^k]$ if $i^k \notin \mathcal{I}^{k-1}$ and $\mathcal{I}^k = \mathcal{I}^{k-1}$ otherwise. Likewise, $\mathcal{J}^k = [\mathcal{J}^{k-1}\ j^k]$ if $j^k \notin \mathcal{J}^{k-1}$ and $\mathcal{J}^k = \mathcal{J}^{k-1}$ otherwise. With the selected indices, the size of C_k will be $|\mathcal{I}^k| \times |\mathcal{J}^k|$ instead of

$k \times k$. To find C_k, it seems natural to solve:

$$C_k = \arg\min_C \frac{1}{2}||\Gamma_{\mathcal{I}^k} C \Psi_{\mathcal{J}^k}^\top - S||_F^2. \tag{3.14}$$

But the solution C_k will contain possible nonzero coefficients that do not coincide with the chosen selection of indices since additional indices in all combinations of pairs between $\mathcal{I}^k$ and $\mathcal{J}^k$ will be updated in each iteration. This is problematic for the correctness of the algorithm when choosing the next single most correlated coefficient. Therefore we must enforce that these coefficients are zero:

$$C_k = \arg\min_C \frac{1}{2}||\Gamma_{\mathcal{I}^k} C \Psi_{\mathcal{J}^k}^\top - S||_F^2 \text{ s.t. } C_{i,j} = 0 \ \forall (i,j) \in \mathcal{O}^k. \tag{3.15}$$

where $\mathcal{O}^k := \overline{(\mathcal{I}^k, \mathcal{J}^k)}$. To solve this problem, we can use projected gradient descent (PGD). The gradient of $f(C) = \frac{1}{2}||\Gamma_{\mathcal{I}^k} C \Psi_{\mathcal{J}^k}^\top - S||_F^2$ at iteration k is

$$\nabla f(C) = \Gamma_{\mathcal{I}^k}^\top \Gamma_{\mathcal{I}^k} C \Psi_{\mathcal{J}^k}^\top \Psi_{\mathcal{J}^k} - \Gamma_{\mathcal{I}^k}^\top S \Psi_{\mathcal{J}^k}. \tag{3.16}$$

Then setting $Z^1 = C_{k-1}$ we iteratively project the update in the gradient direction to the space of feasible solutions:

$$Z^{t+1} = P_{\mathcal{O}^k}(Z^t - \epsilon_k \nabla f(Z^t)), \tag{3.17}$$

where the projection $P_{\mathcal{O}^k}$ sets all elements in $\mathcal{O}^k$ to 0 and the step-size ϵ_k is estimated at each iteration using a line search.

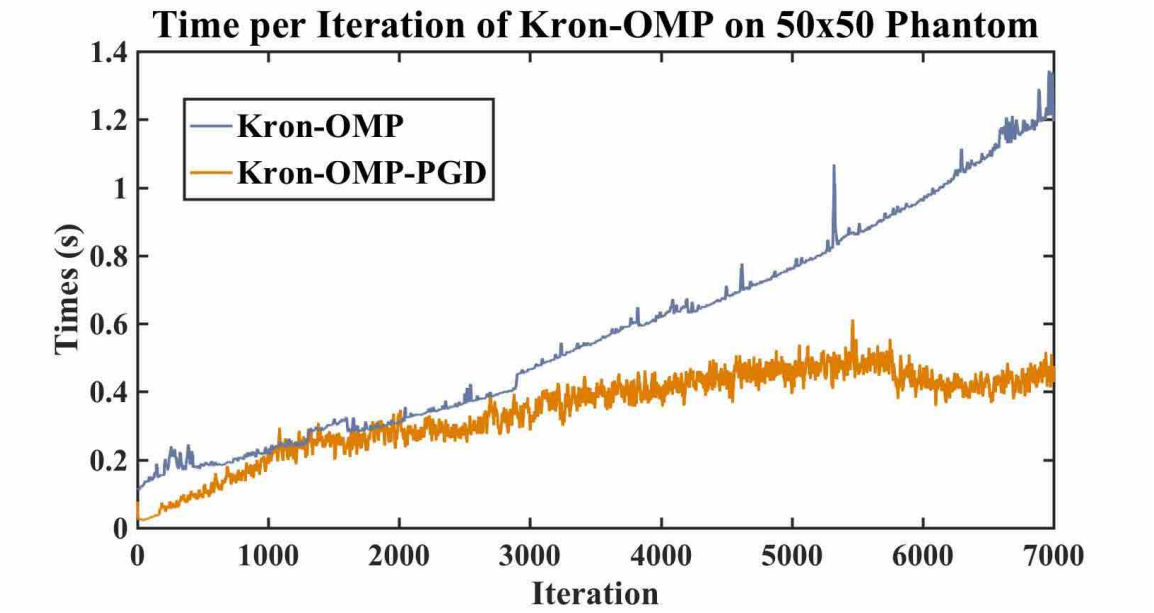


Figure 3.7: Comparison of time per iteration for Kron-OMP and the proposed Kron-OMP-PGD. The total time to choose $K = 7000 = 2.8V$ atoms for this $V = 50 \times 50$ slice of a phantom dataset, is 68 min for Kron-OMP and 40 min for Kron-OMP-PGD. We can see that as the number of atoms grows, the time per iteration of Kron-OMP continues to grow at a much higher rate than Kron-OMP-PGD.

Once the procedure has converged to Z^*, we set $C_k = Z^*$ and compute the residual $R_k = S - \Gamma_{\mathcal{I}^k} C_k \Psi_{\mathcal{J}^k}^\top$. Then, for iteration $k+1$ we must find $(i^{k+1}, j^{k+1}) = \arg\max_{[i,j]} |\Gamma_i^\top R_k \Psi_j|$. To save significantly on computation we precompute $\mathcal{G} = \Gamma^\top \Gamma$, $\mathcal{P} = \Psi^\top \Psi$, and $\hat{S} = \Gamma^\top S \Psi$ and can access access the $\mathcal{I}^k, \mathcal{J}^k$ columns and all rows, at each iteration, using the notation: $\mathcal{G}_{\mathcal{I}^k,\mathcal{I}^k}, \mathcal{P}_{\mathcal{J}^k,\mathcal{J}^k}, \hat{S}_{\mathcal{I}^k,\mathcal{J}^k}$. We use our precomputed $\mathcal{G}$ and $\mathcal{P}$ to instead find $\arg\max_{[i,j]}[\hat{R}_k]_{i,j}$, where $\hat{R}_k = |\hat{S} - \mathcal{G}_{\mathcal{I}^k} C_k \mathcal{P}_{\mathcal{J}^k}^\top|$. In this way, maintaining matrix forms throughout allows us to combine computing the residual and the next atoms for a large reduction in computation at each iteration k. Our pro-

posed Kronecker OMP with projected gradient descent (Kron-OMP-PGD) is outlined in Algorithm 5.

We show a comparison of time per iteration for a small $V = 50 \times 50, G = 64$ phantom dataset in Figure 3.7. The steeper time increase for Kron-OMP is due to the fact that at iteration k there is a complexity of $O(k^2 + kGV)$ that comes from Steps 3 (rank-1 update) and 4 of Algorithm 4. On the other hand, Kron-OMP-PGD has complexity involving $|\mathcal{I}^k|, |\mathcal{J}^k| \leq k$ which are in practice significantly less than k. Even though a PGD sub-routine must be performed at each iteration k, we found that by incorporating Nesterov acceleration with a line search, the time per iteration remains lower than Kron-OMP as the number of iterations k increases.

However, for dMRI data, typical sparsity levels are $K = O(V)$. So for $V \approx 100^3$ the number of iterations as well as the time per iteration of both Kron-OMP and Kron-OMP-PGD when k approaches K becomes astronomical. Even on a relatively small 3D phantom dataset of spatial size $V = 50 \times 50$, for example, one iteration takes on the order of a few seconds which results in over 34 hrs for these greedy algorithms to reach 1 atom/voxel atoms ($K = V$). In this way, greedy algorithms such as OMP are not suitable for large-scale problems that require hundreds of thousands of iterations. Instead, optimizing the LASSO problem ($P1mat$) can be accomplished with significantly less iterations, as we examine in the following section.

Algorithm 5 Kron-OMP-PGD

Choose: $K, \epsilon_1, \epsilon_2$.
Precompute: $\hat{S} = \Gamma^\top S \Psi,\ \mathcal{G} = \Gamma^\top \Gamma,\ \mathcal{P} = \Psi^\top \Psi$.
Initialize: $k = 1,\ C_0 = 1,\ \mathcal{I}^0 = \emptyset,\ \mathcal{J}^0 = \emptyset,\ \hat{R}_0 = \hat{S}$.
while $k \leq K$ and error $> \epsilon_1$ **do**
1: $[i^k, j^k] = \arg\max_{[i,j]} [\hat{R}_k]_{i,j}$;
2: $\mathcal{I}^k = \mathcal{I}^{k-1} \cup \{i^k\}; \mathcal{J}^k = \mathcal{J}^{k-1} \cup \{j^k\}; \mathcal{A}^k = (\mathcal{I}^k, \mathcal{J}^k); \mathcal{O}^k = \overline{\mathcal{A}^k}$;
3: $Z^1_{\mathcal{J}^{k-1}, \mathcal{I}^{k-1}} = C_{k-1};\ n_1 = 0;\ t = 1$;
while error $> \epsilon_2$ **do**
1: $\delta = \text{linesearch}(Z^t)$;
2: $X^{t+1} = P_{\mathcal{O}^k}(Z^t - \delta(\mathcal{G}_{\mathcal{I}^k, \mathcal{I}^k} Z^t \mathcal{P}_{\mathcal{J}^k, \mathcal{J}^k} - \hat{S}_{\mathcal{I}^k, \mathcal{J}^k}))$;
4: $n_{t+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_t^2})$;
5: $Z^{t+1} = X^{t+1} + \frac{n_t - 1}{n_{t+1}}(X^{t+1} - X^t)$;
6: $t \leftarrow t + 1$;
end while
4: $C_k = Z^*$;
5: $\hat{R}_k = |\hat{S} - \mathcal{G}_{\mathcal{I}^k} C_k \mathcal{P}_{\mathcal{J}^k}|$;
6: $k \leftarrow k + 1$;
end while
Return: $\mathcal{A}^K, C_K$.

3.4.3 Kronecker ADMM

The Alternating Direction Method of Multipliers (ADMM) [102] is a popular method for solving the LASSO problem ($P1vec$). However, its application in the case of a large dictionary Φ remains prohibitive, requiring computations involving $\Phi^\top s$ of order $O(GVN_\Gamma N_\Psi)$. Instead, we apply ADMM to the separable L_1 minimization problem ($P1mat$) to reduce computations by solving

$$\min_{C,Z} \frac{1}{2} ||\Gamma C \Psi^\top - S||_F^2 + \lambda ||Z||_1 \ \text{ s.t. } \ C = Z. \tag{3.18}$$

The augmented Lagrangian writes:

$$\mathcal{L}_\mu(C, Z, \mathcal{T}) = \frac{1}{2}||\Gamma C \Psi^\top - S||_F^2 + \lambda||Z||_1 + < \mathcal{T}, C - Z > + \frac{\mu}{2}||C - Z||_F^2, \quad (3.19)$$

and:

$$\frac{\partial \mathcal{L}_\mu}{\partial C} = \Gamma^\top(\Gamma C \Psi^\top - S)\Psi + \mathcal{T} + \mu(C - Z) = 0 \quad (3.20)$$

$$\implies \Gamma^\top \Gamma C \Psi^\top \Psi + \mu C = \mu Z - \mathcal{T} + \Gamma^\top S \Psi := Q. \quad (3.21)$$

In principle, one can solve for C by solving a linear system of equations $h(C) = Q$, where $h(C) = \Gamma^\top \Gamma C \Psi^\top \Psi + \mu C$. However, solving this linear system directly is computationally challenging due to the size of the matrices involved. Therefore, to solve for C efficiently, we begin by taking the SVDs of Γ and Ψ. With $\Gamma = U_\Gamma \Sigma_\Gamma V_\Gamma^\top$ and $\Psi = U_\Psi \Sigma_\Psi V_\Psi^\top$, $\Gamma^\top \Gamma = V_\Gamma \Delta_\Gamma V_\Gamma^\top$ and $\Psi^\top \Psi = V_\Psi \Delta_\Psi V_\Psi^\top$, where U_Γ, U_Ψ are the matrices of eigenvectors and $\Delta_\Gamma = \Sigma_\Gamma^\top \Sigma_\Gamma, \Delta_\Psi = \Sigma_\Psi^\top \Sigma_\Psi$ are the diagonal matrices of eigenvalues for Γ and Ψ respectively. Then:

$$V_\Gamma \Delta_\Gamma V_\Gamma^\top C V_\Psi \Delta_\Psi V_\Psi^\top + \mu C = Q \quad (3.22)$$

$$\implies \Delta_\Gamma \tilde{C} \Delta_\Psi + \mu \tilde{C} = \tilde{Q} \quad (3.23)$$

where we introduced the notation $\tilde{X} = V_\Gamma^\top X V_\Psi$. Since Δ_Γ and Δ_Ψ are diagonal with

elements δ_{Γ_i} and δ_{Ψ_j}, respectively, we can solve for $\tilde{C}$ by:

$$\delta_{\Gamma_i}\tilde{C}_{i,j}\delta_{\Psi_j} + \mu\tilde{C}_{i,j} = \tilde{Q}_{i,j} \implies \tilde{C}_{i,j} = \frac{\tilde{Q}_{i,j}}{\delta_{\Gamma_i}\delta_{\Psi_j} + \mu}. \tag{3.24}$$

To write this in matrix form we define $[\Delta_\mu]_{i,j} \triangleq 1/(\delta_{\Gamma_i}\delta_{\Psi_j} + \mu)$ and have $\tilde{C} = (\Delta_\mu \circ \tilde{Q})$ where $\circ$ stands for element-wise matrix multiplication. Finally, we can recover $C = V_\Gamma \tilde{C} V_\Psi^\top$ and the complete update for C is:

$$C_{k+1} = V_\Gamma(\Delta_\mu \circ (V_\Gamma^\top Q_k V_\Psi))V_\Psi^\top \tag{3.25}$$

where $Q = \mu Z - \mathcal{T} + \Gamma^\top S \Psi$. When minimizing $\mathcal{L}_\mu$ with respect to Z, we end up with the usual proximal operator of the L_1 norm that is given by the shrinkage operator, $\text{shrink}_\kappa(X) = \max(0, X - \kappa) - \max(0, -X - \kappa)$, applied element-wise to matrix X, giving $Z_{k+1} = \text{shrink}_{\lambda/\mu}(C_{k+1} + \mathcal{T}_k)$. Similarly with respect to $\mathcal{T}$, we have the usual Lagrange multiplier gradient ascent update $\mathcal{T}_{k+1} = \mathcal{T}_k + C_{k+1} - Z_{k+1}$. The formal updates for Kron-ADMM are presented in Algorithm 6. The update for C in (3.25) works well when Γ and Ψ are under-complete and the eigen-decompositions of $\Gamma^\top\Gamma$ and $\Psi^\top\Psi$ are easily computable. However, dictionaries most commonly used for sparse coding and the application to CS are over-complete i.e. $G < N_\Gamma$ and $V < N_\Psi$ making these SVDs potentially expensive to compute. In the case of an over-complete Φ, for traditional vector ADMM, the matrix inversion lemma [102] is involved in order to compute SVDs of the smaller $\Phi\Phi^\top$ instead of $\Phi^\top\Phi$. In the following proposition,

Algorithm 6 Kron-ADMM (for undercomplete dictionaries)

Choose: μ, λ, ϵ.
Precompute: $V_\Gamma, V_\Psi, \Delta_\mu$.
Initialize: $k = 0, Z_0 = \mathbf{0}, \mathcal{T}_0 = \mathbf{0}$.
while error $> \epsilon$ **do**
1: $Q_k = \Gamma^\top S\Psi + \mu Z_k - \mathcal{T}_k$;
2: $C_{k+1} = V_\Gamma(\Delta_\mu \circ (V_\Gamma^\top Q_k V_\Psi))V_\Psi^\top$;
3: $Z_{k+1} = \text{shrink}_{\lambda/\mu}(C_{k+1} + \mathcal{T}_k)$;
4: $\mathcal{T}_{k+1} = \mathcal{T}_k + C_{k+1} - Z_{k+1}$;
5: $k \leftarrow k + 1$;
end while
Return: C.

we derive the equivalent result for the update of C in (3.25).

Proposition 1. *For over-complete dictionaries* Γ *and* Ψ*, update* (3.25) *is equivalent to the more compact*

$$C = Q/\mu - \Gamma^\top U_\Gamma(\Delta_\mu \circ (U_\Gamma^\top \Gamma Q \Psi^\top U_\Psi))U_\Psi^\top \Psi/\mu. \tag{3.26}$$

Proof. For over-complete dictionaries $\Gamma = U_\Gamma[\Sigma_\Gamma,\ \mathbf{0}]V_\Gamma^\top$ and $\Psi = U_\Psi[\Sigma_\Psi,\ \mathbf{0}]V_\Psi^\top$,

$$\Gamma^\top\Gamma = V_\Gamma \begin{pmatrix} \Delta_\Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} V_\Gamma^\top \text{ and } \Psi^\top\Psi = V_\Psi \begin{pmatrix} \Delta_\Psi & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} V_\Psi^\top.$$

For $G < i \leq N_\Gamma, V < j \leq N_\Psi$, $\delta_{\Gamma_i}, \delta_{\Psi_j} = 0$, so $\tilde{C}_{i,j} = \frac{\tilde{Q}_{i,j}}{\delta_{\Gamma_i}\delta_{\Psi_j}+\mu} = \frac{\tilde{Q}_{i,j}}{\mu}$. For $i \leq G$ and

$j \leq V$, we can rewrite

$$\begin{aligned}
\tilde{C}_{i,j} = \frac{\tilde{Q}_{i,j}}{\delta_{\Gamma_i}\delta_{\Psi_j} + \mu} &= \frac{\tilde{Q}_{i,j}}{\mu} - \frac{\delta_{\Gamma_i}\tilde{Q}_{i,j}\delta_{\Psi_j}}{\mu(\delta_{\Gamma_i}\delta_{\Psi_j} + \mu)} = \frac{\tilde{Q}_{i,j}}{\mu} - \frac{\sigma^2_{\Gamma_i}\tilde{Q}_{i,j}\sigma^2_{\Psi_j}}{\mu(\delta_{\Gamma_i}\delta_{\Psi_j} + \mu)} \\
&= \frac{\tilde{Q}_{i,j}}{\mu} - \sigma_{\Gamma_i}\frac{\sigma_{\Gamma_i}\tilde{Q}_{i,j}\sigma_{\Psi_j}}{\mu(\delta_{\Gamma_i}\delta_{\Psi_j} + \mu)}\sigma_{\Psi_j} \\
\implies \tilde{C} &= \tilde{Q}/\mu - \Sigma_\Gamma^\top(\Delta_\mu \circ (\Sigma_\Gamma \tilde{Q} \Sigma_\Psi^\top))\Sigma_\Psi/\mu \\
C &= Q/\mu - V_\Gamma \Sigma_\Gamma^\top(\Delta_\mu \circ (\Sigma_\Gamma V_\Gamma^\top Q V_\Psi \Sigma_\Psi^\top))\Sigma_\Psi V_\Psi^\top/\mu \\
C &= Q/\mu - \Gamma^\top U_\Gamma(\Delta_\mu \circ (U_\Gamma^\top \Gamma Q \Psi^\top U_\Psi))U_\Psi^\top \Psi/\mu
\end{aligned}$$

□

Letting $\Gamma' = U_\Gamma^\top \Gamma$ and $\Psi' = U_\Psi^\top \Psi$, which can be precomputed, we have a final efficient update

$$C_{k+1} = Q_k/\mu - \Gamma'^\top(\Delta_\mu \circ (\Gamma' Q_k \Psi'^\top))\Psi'/\mu \tag{3.27}$$

This allows us to compute the SVDs of $\Gamma\Gamma^\top$ and $\Psi\Psi^\top$ instead of the larger $\Gamma^\top\Gamma$ and $\Psi^\top\Psi$ and work with smaller matrices within each iteration. We present Kron-ADMM for over-complete dictionaries in Algorithm 7.

3.4.4 Kronecker Dual ADMM

As an alternative to ADMM, Dual ADMM, which applies ADMM to the dual of L_1 problem ($P1vec$), has been shown to be more efficient than ADMM for over-complete dictionaries [134] by allowing one to compute SVDs of the more affordable $\Phi\Phi^\top$

Algorithm 7 Kron-ADMM (for overcomplete dictionaries)

Choose: μ, λ, ϵ.
Precompute: $U_\Gamma, \Delta_\Gamma, U_\Psi, \Delta_\Psi, \Gamma', \Psi', \Delta_\mu$.
Initialize: $k = 0, Z_0 = \mathbf{0}, \mathcal{T}_0 = \mathbf{0}$.
while error $> \epsilon$ **do**
1: $Q_k = \Gamma^\top S\Psi + \mu Z_k - \mathcal{T}_k$;
2: $C_{k+1} = Q_k/\mu - \Gamma'^\top(\Delta_\mu \circ (\Gamma' Q_k \Psi'^\top))\Psi'/\mu$;
3: $Z_{k+1} = \text{shrink}_{\lambda/\mu}(C_{k+1} + \mathcal{T}_k)$;
4: $\mathcal{T}_{k+1} = \mathcal{T}_k + C_{k+1} - Z_{k+1}$;
5: $k \leftarrow k + 1$;
end while
Return: C.

instead of $\Phi^\top\Phi$. In our previous work [84] we proposed a Kronecker Dual ADMM (Kron-DADMM) that efficiently solves the spatial-angular sparse coding problem. Below, we give an alternative derivation of this algorithm directly based on the matrix formulation of $(P1mat)$. The dual of $(P1mat)$ is:

$$\max_A -\frac{1}{2}||A||_F^2 + A^\top S \quad \text{s.t.} \quad ||\Gamma^\top A\Psi||_\infty \leq \lambda, \tag{3.28}$$

where $||X||_\infty = \max_{i,j} |X_{i,j}|$. To apply ADMM to this optimization problem, we replace $\Gamma^\top A\Psi$ with auxiliary variable $\mathcal{V}$ and add the additional constraint $\Gamma^\top A\Psi - \mathcal{V} = 0$ to get:

$$\max_{A,\mathcal{V}} -\frac{1}{2}||A||_F^2 + A^\top S \quad \text{s.t.} \quad ||\mathcal{V}||_\infty \leq \lambda \quad \text{and} \quad \mathcal{V} = \Gamma^\top A\Psi. \tag{3.29}$$

Then the augmented Lagrangian is

$$\mathcal{L}_\eta(A, \mathcal{V}, C) = -\frac{1}{2}||A||_F^2 + A^\top S + < C, \mathcal{V} - \Gamma^\top A\Psi > + \frac{\eta}{2}||\mathcal{V} - \Gamma^\top A\Psi||_F^2 + \delta_\lambda(\mathcal{V}) \tag{3.30}$$

where

$$\delta_\lambda(\mathcal{V}) = \begin{cases} 0 & \text{if } ||\mathcal{V}||_\infty \leq \lambda \\ \\ \infty & \text{if } ||\mathcal{V}||_\infty > \lambda \end{cases}. \tag{3.31}$$

and the Lagrange multiplier C corresponds to the primal variable C in $(P1mat)$, which our variable of interest. We then have

$$\frac{\partial \mathcal{L}_\eta(A, \mathcal{V}, C)}{\partial A} = -A + S - \Gamma C \Psi^\top - \eta\Gamma(\mathcal{V} - \Gamma^\top A \Psi)\Psi^\top = 0 \tag{3.32}$$

$$\implies A - \eta\Gamma\Gamma^\top A \Psi\Psi^\top = S - \Gamma(C + \eta\mathcal{V})\Psi^\top := P. \tag{3.33}$$

Now with eigen-decompositions $\Gamma\Gamma^\top = U_\Gamma \Delta_\Gamma U_\Gamma^\top$ and $\Psi\Psi^\top = U_\Psi \Delta_\Psi U_\Psi^\top$ and letting $\tilde{X} = U_\Gamma^\top X U_\Psi$,

$$A + \eta U_\Gamma \Delta_\Gamma U_\Gamma^\top A U_\Psi \Delta_\Psi U_\Psi^\top = P \tag{3.34}$$

$$\implies \tilde{A} + \eta\Delta_\Gamma \tilde{A} \Delta_\Psi = \tilde{P}. \tag{3.35}$$

Then, $\tilde{A}$ can be found element-wise by:

$$\tilde{A}_{i,j} + \eta\delta_{\Gamma_i}\tilde{A}_{i,j}\delta_{\Psi_j} = \tilde{P}_{i,j} \implies \tilde{A}_{i,j} = \frac{\tilde{P}_{i,j}}{1 + \eta\delta_{\Gamma_i}\delta_{\Psi_j}}. \tag{3.36}$$

Defining $[\Delta_\eta]_{i,j} \triangleq 1/(1 + \eta\delta_{\Gamma_i}\delta_{\Psi_j})$, the update is $\tilde{A} = \Delta_\eta \circ \tilde{P}$. As shown in [134] we can keep the update in terms of $\tilde{A}$ instead of A since the variable we are interested in is C. We can then precompute $S' = \Gamma'^\top S \Psi'$, $\Gamma' = U_\Gamma^\top \Gamma$ and $\Psi' = U_\Psi^\top \Psi$. The updates

of $\mathcal{V}$ and C are as in [84] and presented in Algorithm 8, where $P_\lambda^\infty(X)$ sets all entries of matrix X that are greater than λ to λ.

Algorithm 8 Kron-DADMM

Choose: η, λ, ϵ.
Precompute: $S', \Gamma', \Psi', \Delta_\eta$.
Initialize: $k = 0, C_0 = 0, \mathcal{V}_0 = 0$.
while Duality Gap $> \epsilon$ **do**
1: $\tilde{A}_{k+1} = \Delta_\eta \circ (S' - \Gamma'(C_k - \eta\mathcal{V}_k)\Psi'^\top)$;
2: $\mathcal{V}_{k+1} = P_\lambda^\infty(\frac{1}{\eta}C_k + \Gamma'^\top \tilde{A}_{k+1}\Psi')$;
3: $C_{k+1} = \text{shrink}_{\lambda\eta}(C_k + \eta\Gamma'^\top \tilde{A}_{k+1}\Psi')$;
4: $k \leftarrow k + 1$;
end while
Return: C.

3.4.5 Kronecker FISTA

The Fast Iterative Thresholding Algorithm (FISTA) [103] is another well-known method for solving LASSO. However, just as before, applying FISTA to ($P1vec$) for large-scale dMRI data is largely intractable. So here we adapt FISTA to ($P1mat$) in order to exploit the separability of our spatial-angular basis. FISTA is a proximal gradient descent

$$C_{k+1} = \text{shrink}_{\lambda/L}(C_k - \nabla f(C_k)/L), \tag{3.37}$$

where the proximal operator is the soft-thresholding shrinkage operator associated with the L_1 norm and $1/L$ is a chosen step size. The gradient is simply computed as:

$$\nabla f(C) = \Gamma^\top(\Gamma C \Psi^\top)\Psi - \Gamma^\top S\Psi. \tag{3.38}$$

To help speed convergence, we use a line search subroutine to update L at each iteration in addition to the usual Nesterov acceleration. By [103], FISTA will converge for any L greater than the Lipschitz constant of ∇f, which can be estimated by bounding

$$||\nabla f(C) - \nabla f(\bar{C})||_F = ||\Gamma^\top \Gamma (C - \bar{C}) \Psi^\top \Psi||_F \leq \lambda^\Gamma_{\max} \lambda^\Psi_{\max} ||C - \bar{C}||_F \tag{3.39}$$

where $\lambda^\Gamma_{\max}$ and $\lambda^\Psi_{\max}$ are the maximum eigenvalues of $\Gamma^\top\Gamma$ and $\Psi^\top\Psi$ respectively. Therefore we initialize $L = \lambda^\Gamma_{\max}\lambda^\Psi_{\max}$. The Kronecker FISTA (Kron-FISTA) is presented in Algorithm 9. This natural Kronecker extension to FISTA has also been recently presented in [135], but has not been adapted and tested on data of our scale.

Algorithm 9 Kron-FISTA

Choose: ϵ.
Precompute: $\hat{S} = \Gamma^\top S \Psi$
Initialize: $Z_1 = C_0 = \mathbf{0}$, $n_1 = 1, L = \lambda^\Gamma_{\max}\lambda^\Psi_{\max}$.
while error $> \epsilon$ **do**
1: $L = \text{linesearch}(Z_k)$;
2: $\nabla f(Z_k) = \Gamma^\top(\Gamma Z_k \Psi^\top)\Psi - \hat{S}$;
3: $C_k = \text{shrink}_{\lambda/L}(Z_k - \nabla f(Z_k)/L)$;
4: $n_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_k^2})$;
5: $Z_{k+1} = C_{k+1} + \frac{n_k - 1}{n_{k+1}}(C_{k+1} - C_k)$;
6: $k \leftarrow k + 1$;
end while
Return: C.

Algorithm	Standard	Kronecker
OMP	$k^2 + kGV + GVN_\Gamma N_\Psi$	$k^2 + kGV + GVN_\Gamma + VN_\Gamma N_\Psi$
OMP-PGD	–	$TG\|\mathcal{I}^k\|\|\mathcal{J}^k\| + TGV\|\mathcal{J}^k\| + \|\mathcal{J}^k\|N_\Gamma N_\Psi$
ADMM	$(GV)^2 N_\Gamma N_\Psi + GV(N_\Gamma N_\Psi)^2$	$(GN_\Gamma N_\Psi + GVN_\Psi) + GV$
DADMM	$(GV)^2 N_\Gamma N_\Psi$	$(GN_\Gamma N_\Psi + GVN_\Psi) + GV$
FISTA	$(N_\Gamma N_\Psi)^2 + GVN_\Gamma N_\Psi$	$(GN_\Gamma N_\Psi + GVN_\Psi)$

Table 3.3: Comparison of algorithms complexity at iteration k. For Kron-OMP-PGD, T is the number of sub-iterations of PGD.

3.4.6 Complexity Analysis

To evaluate the efficiency of each algorithm and the gains of Kronecker separability compared to the original algorithms we summarize the complexity of each algorithm for general Ψ and Γ in Table 3.3. We notice that classical L_1 algorithms have complexity on the order of the size of the Φ matrix, including terms that multiply all four dimensions $GVN_\Gamma N_\Psi$. When applying the Kronecker L_1 algorithms, the complexity is reduced to a summation that includes only 3 of the dimensions GVN_Ψ, a reduction on the order of N_Γ (≈ 200 for some of our dictionary choices). We compare the Kronecker L_1 algorithms empirically in Section 3.5 to identify which is fastest for our regime. Next we address the fact that the dimensions of $\Gamma \in \mathbb{R}^{G\times N_\Gamma}$ and $\Psi \in \mathbb{R}^{V\times N_\Psi}$ will be orders of magnitude different since $G \approx 100$ and $V \approx 100^3$. We consider a few specific assumptions on the structure of spatial dictionary Ψ which can decrease the complexity and simplify computations of some of the proposed algorithms:

Ψ Tight Frame. In the case that Ψ is a tight frame, $\Psi^\top = \mathrm{I}$, which is commonly an assumption in compressed sensing theorems, our method can still be simplified. In Kron-ADMM (overcomplete) and Kron-DADMM, we may avoid the SVD of $\Psi\Psi^\top$

and respective updates (3.23) and (3.35) can be simplified.

Ψ Fast Transform. In the case that Ψ corresponds to a well-studied transform such as wavelets, curvelets, etc., fast transform implementations can be utilized to reduce complexity further. For the case of FISTA, for example, matrix multiplications of $\Gamma^\top(\Gamma Z_k \Psi^\top)\Psi$ (See Algorithm 9 Step 2) involve fast transform reconstructions ($\Psi^\top$) of each DWI (ΓZ_k) and then deconstructions (Ψ) which we parallelize over all DWI in our implementation.

Ψ Orthonormal. In the case that Ψ is orthonormal, ${}^\top\Psi = \Psi\Psi^\top = \mathrm{I}$ then ($P1mat$) can be simplified to (3.5) after noticing:

$$||\Gamma C \Psi^\top - S||_F^2 = ||\Gamma C \Psi^\top \Psi - S\Psi||_F^2 = ||\Gamma C - \hat{S}||_F^2. \tag{3.40}$$

This optimization can be solved using traditional methods after precomputing $\hat{S} = S\Psi$.

Ψ Separable Tensor Product. In the case that Ψ can be separated into a 3D tensor product $\Psi = \Psi_x \otimes \Psi_y \otimes \Psi_z$, the complexity of multiplication can be simplified by another degree, in the same vein as the decrease in complexity we gained from using $\Phi = \Psi \otimes \Gamma$. In this case, instead of the matrix multiplication, $S = \Gamma C \Psi^\top$ can be written using n-mode products of tensors $\mathcal{S} = \mathcal{C} \times_x \Psi_x \times_y \Psi_y \times_z \Psi_z \times_q \Gamma$. Furthermore, if we consider DSI acquisition where q-space measurements are acquired in a grid over $\mathbb{R}^3$, and assume we can represent these measurements over a separable basis over each

dimension, then we can take $\Gamma = \Gamma_{q_x} \otimes \Gamma_{q_y} \otimes \Gamma_{q_z}$ and Φ becomes a 6-tensor.

3.5 Experiments on Spatial-Angular Sparse Coding

3.5.1 Data

We perform our experiments on single-shell HARDI data, though as we emphasized earlier, our framework and algorithms can be applied to any dMRI acquisition protocol with a suitable choice of the angular basis Γ. Specifically, we experimented on a phantom and a real HARDI brain dataset. We applied our methods to the ISBI 2013 HARDI Reconstruction Challenge Phantom dataset[3], a $V = 50 \times 50 \times 50$ volume consisting of 20 phantom fibers crossing intricately within an inscribed sphere, measured with $G = 64$ gradient directions (SNR = 30 dB). Our initial experiments test on a 2D 50×50 slice of this data for simplification. The real HARDI brain dataset consists of a $V = 112 \times 112 \times 65$ volume with $G = 127$ gradient directions. We conducted experiments on the core white matter brain region of size $V = 60 \times 60 \times 30$.

[3] http://hardi.epfl.ch/static/events/2013_ISBI/

Atoms/Voxel	0.09	0.24	0.60	1.72	3.67	6.75
Lambda	1.4^{1}	1.4^{-1}	1.4^{-3}	1.4^{-5}	1.4^{-7}	1.4^{-9}
Kron-ADMM	797	1462	2096	3660	4365	4667
Kron-DADMM	357	597	1060	1722	1928	1953
Kron-FISTA	161	219	288	346	584	611

Table 3.4: Number of iterations to completion for Kron-ADMM, Kron-DADMM, Kron-FISTA. For computation time, see Figure 3.8.

3.5.2 Kronecker Algorithm Comparison

In this section we compare the computational time performance of each of the proposed Kronecker LASSO algorithms, Kron-ADMM, Kron-DADMM, and Kron-FISTA on a 2D 50×50 slice of phantom data for various values of λ using Haar-SR. For our experiment, we ran Kron-FISTA until a very small error of 10^{-8} was reached. The objective value obtained was then taken to be a rough ground truth minimum. We then tested each of Kron-ADMM, Kron-DADMM, and Kron-FISTA and recorded the time it took to reach a relative error of 10^{-4} from the known minimum. Figure 3.8 reports the objective value descent of each algorithm for various sparsity levels associated to choices of λ. Table 3.8 gives the number of iterations until completion for each method and sparsity level. For our experiments, Kron-FISTA appears to be the fastest algorithm in all cases, followed by Kron-DADMM. The superior performance of DADMM over ADMM is consistent with the findings of [134]. With these results, we henceforth use Kron-FISTA for subsequent experiments.

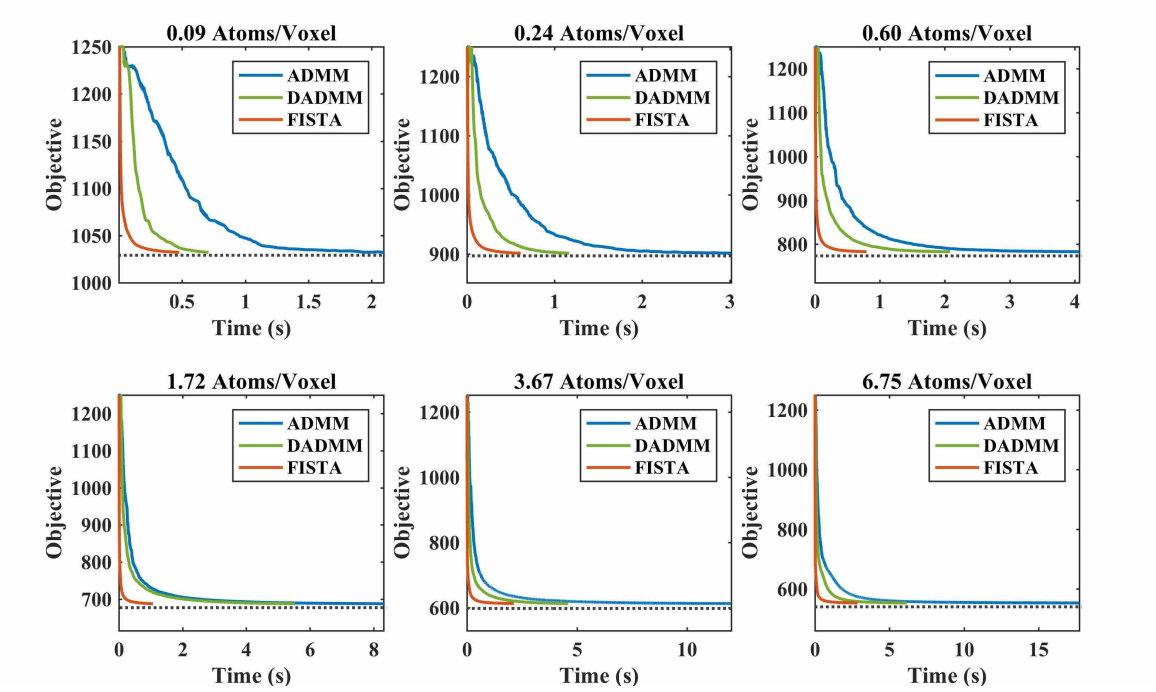


Figure 3.8: Comparison of time for completion of Kron-ADMM, Kron-DADMM, and Kron-FISTA on a 2D 50×50 phantom HARDI data using Haar-SR for various sparsity levels. Kron-FISTA consistently reaches the known minimum objective in the least amount of time. For number of iterations and lambda values, see Table 3.4.

3.5.3 Choice of Spatial-Angular Dictionaries

The experiments in this chapter are conducted using fixed spatial and angular dictionaries. The best performing dictionaries will be used throughout this thesis when fixed dictionaries are used.

Choice of Spatial Dictionary. For the choice of spatial dictionary Ψ, the spatial wavelet transform is a popular sparsifying basis for natural images and structural MRI. The simplest wavelet transform is Haar, square-shaped functions that provide sharp transitions from positive to negative. Alternatively, Daubechies wavelets provide a smoother boundaries and the higher the order $(2, 3, 4, \dots)$, the more complex

signals it can model efficiently. For notation we will write Db4 for Daubechies wavelets of order 4, for example.

In addition to wavelets, we consider the spatial curvelets dictionary [136] (featured as the spatial atom in Figure 3.5) which, in addition to variations in position and scale, offers directional variations which may be useful for sparsely modeling the naturally directional HARDI fiber tracts regions. An important criteria for choosing our spatial basis is that they be tight frames as this choice has important theoretical implications for compressed sensing (as we will see in Chapter 4) and offers computational advantages (as discussed in Section 3.4.6). They additionally have fast transform implementations which also reduce computational complexity.

Finally, to compare our formulation to state-of-the-art voxel-wise angular sparse coding, we can simply choose Ψ to be the $V \times V$ identity I_V (refer to the end of Section 3.3 for the derivation).

Choice of Angular Dictionary. For the choice of the angular dictionary, one well-known choice is the spherical harmonics (SH) basis defined in Chapter 2.1.3.3. Like the Fourier basis, the SH basis is useful for reconstruction of any bandlimited spherical function. For order $L = 4$, for example, the SH basis has $N_\Gamma = 15$ atoms, which has been shown to be enough atoms to accurately reconstruct an q-space signal. Since SH is undercomplete at this order, it has been shown to be ill-suited for sparse coding [44], producin. In our experiments we aim to use on the order of 1 atom/voxel instead of 15.

As an alternative, the over-complete Spherical Ridgelet (SR) dictionary [44], derived in Chapter 2.1.3.3 has been shown to sparsely model HARDI signals. The corresponding dictionary in the space of ODFs is the set of spherical wavelets (SW) (see Figure 3.5 for an example of one spherical wavelet atom). With order $L = 2$ and 4, the SR dictionary contains $N_\Gamma = 210$ and $N_\Gamma = 1169$ atoms, respectively. We used both amounts of atoms for the small 2D 50×50 phantom dataset and found roughly identical results suggesting that a basis of order $L = 2$ contains enough atoms if the number of gradients is below 210. This reduces computation significantly.

For ease of notation, we use a spatial-angular Ψ-Γ labeling: Haar-SR, Db-SR, Curve-SR, I-SR for Haar wavelets, Daubechies wavelets, curvelets, and the identity, respectively, for the spatial domain with SR for the angular domain.

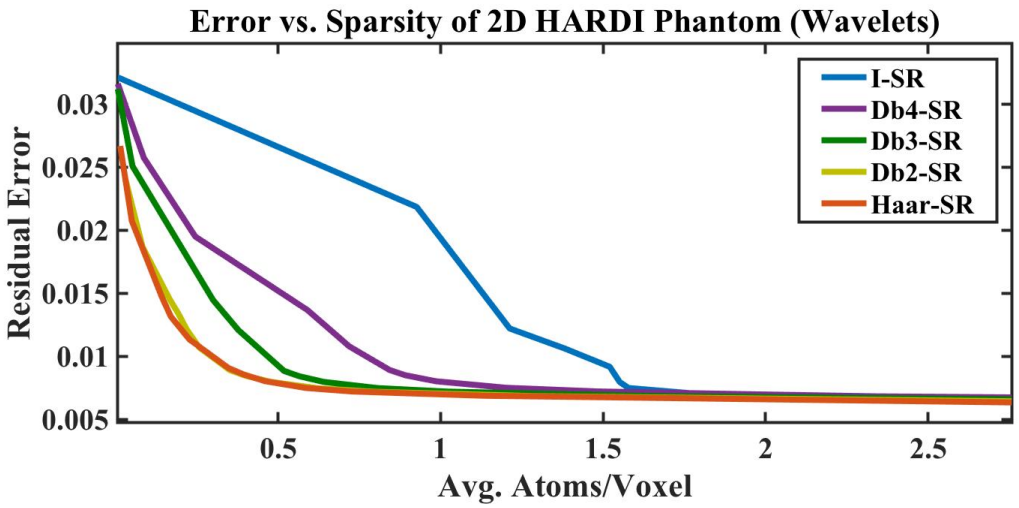


Figure 3.9: Quantitative results of residual error vs. spatial-angular sparsity levels for I-SR, Db4-SR, Db3-SR, Db2-SR, and Haar-SR, on 2D phantom data for various values of λ. Haar wavelets outperforms Daubechies wavelets of all orders. I-SR has a higher reconstruction error at sparsity levels less than 1 atom/voxel.

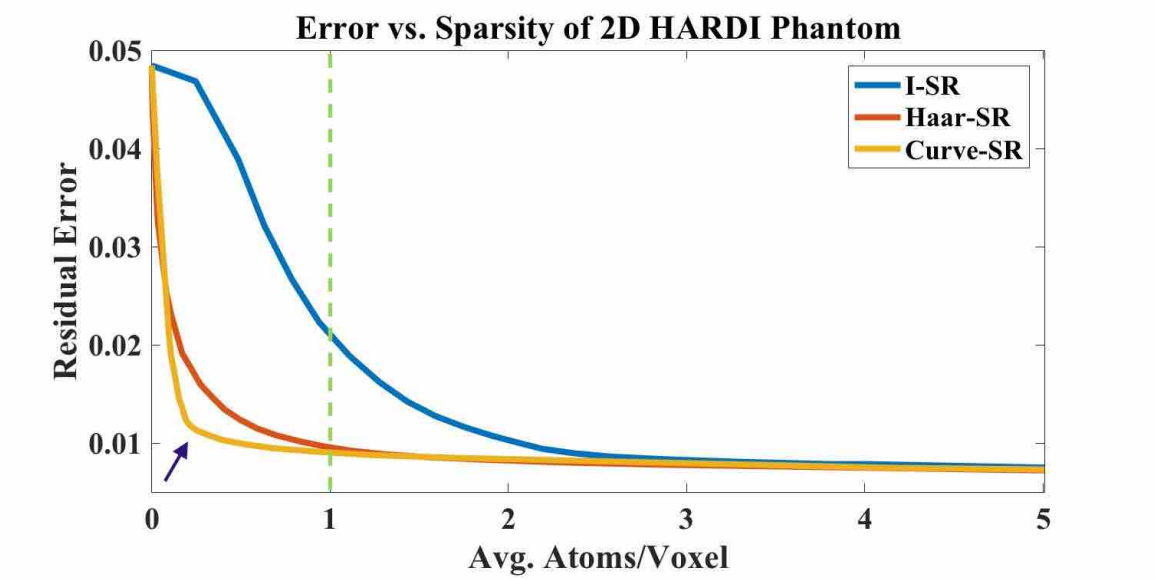


Figure 3.10: Quantitative results of residual error vs. spatial-angular sparsity levels for I-SR, Haar-SR, and Curve-SR on 2D phantom data for various values of λ. Curve-SR out performs Haar-SR while I-SR has very high relative reconstruction error. The reconstruction of I-SR data points are displayed in Figure 3.2 and Haar-SR/Curve-SR in Figure 3.11. Our finding of I-SR requiring 6-8 atoms per voxel for accurate reconstruction is consistent with previous findings [14, 15].

3.5.4 Sparsity Results

In this section we compare the performance of our spatial-angular sparse coding method to the state-of-the-art angular sparse coding by analyzing reconstruction accuracy using very few nonzero coefficients. Will show qualitative results of ODFs to visualize the impact of sparsity and accompanying quantitative results of residual reconstruction error $\frac{1}{GV}||S^* - S_{orig}||_F$ vs. spatial-angular sparsity levels in terms of the average number of atoms per voxel ($||C^*||_0/V$). The ideal reconstruction will have a very low average number of atoms per voxel with low residual error, which

coincides with the lower left-hand corner of each quantitative plot.

The first experiment is tested on a 50×50 phantom HARDI data slice to compare the performance of the different choices of wavelets, Haar and Daubechies order 2, 3 and 4, with the state-of-the-art voxel based sparse coding using the identity (I-SR). This particular experiment was run using Kron-DADMM for various values of λ and featured in our previous work [84]. Figure 3.9 demonstrates that Haar wavelets outperforms Daubechies wavelets of all orders in terms of reconstruction error arguably due to the fact that HARDI data exhibits more rigid boundaries and piece-wise consistencies between isotropic and anisotropic regions which are smoothed by Daubechies wavelets. The wavelets all out-perform the state of the art.

For the next experiment we ran Kron-FISTA for various values of λ using the best performing wavelet basis, Haar-SR, with Curve-SR and I-SR. In Figure 3.10 we show the results of residual reconstruction error vs. spatial-angular sparsity. We can see that in this range, Curve-SR outperforms Haar-SR while I-SR is unable to perform at this level. Reconstruction of I-SR for various sparsity levels are visualized in Figure 3.2. In comparison, Figure 3.11 displays the sparse reconstruction of Haar-SR and Curve-SR with an average of 0.25 atoms/voxel. Notice that Curve-SR leads to a somehow smoother and more accurate reconstruction than the expectedly boxy reconstruction of Haar-SR at this very high sparsity level. Still, in both cases, the proposed joint spatial-angular sparse coding can reconstruct accurate signals with much fewer number of atoms than angular sparse coding, which as seen again from

Figure 3.2 can be achieved with an average of around 4 atoms per voxel. More strikingly, in cases of high signal complexity for crossing fibers, the sparse code requires on the order of 6-12 atoms per voxel (see Figure 3.4).

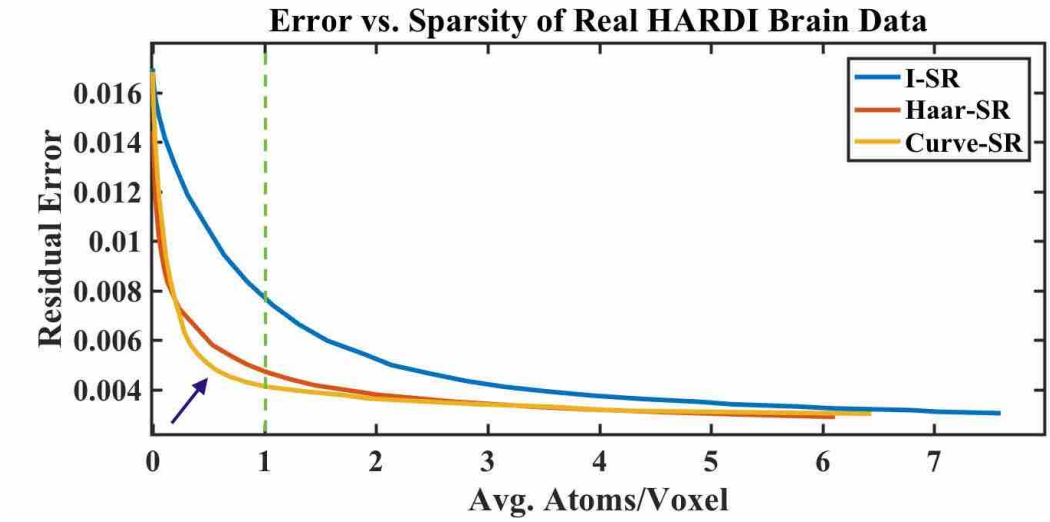


Figure 3.12: Comparison of the spatial-angular sparsity level achieved by Haar-SR and Curve-SR with respect to the state-of-the-art I-SR. The curvelets provide a good reconstruction error with the sparsest number of atoms, in the range of 0.5 to 2 atoms/voxel. The state-of-the-art error is much larger in this sparsity range and only comparable in the predicted range of 6-8 atoms/voxel, consistent with the previously reported [14,15] for I-SR.

We repeated this same analysis on real HARDI data. Figure 3.12 presents the reconstruction error vs. sparsity results for I-SR, Haar-SR, and Curve-SR showing again that curvelets outperforms Haar for high sparsity levels in the range of 0.5-2 avg. atoms/voxel. As expected and consistent with our phantom data experiment, the state-of-the-art I-SR has comparable reconstruction error in the range of 6-8 avg. atoms/voxel. Figure 3.13 shows the quality of reconstruction of I-SR, Haar-SR, and Curve-SR compared to the original signal for the high sparsity level of $\sim$ 1 avg. atom/voxel. Haar-SR presents boxy regions while Curve-SR maintains a smoother

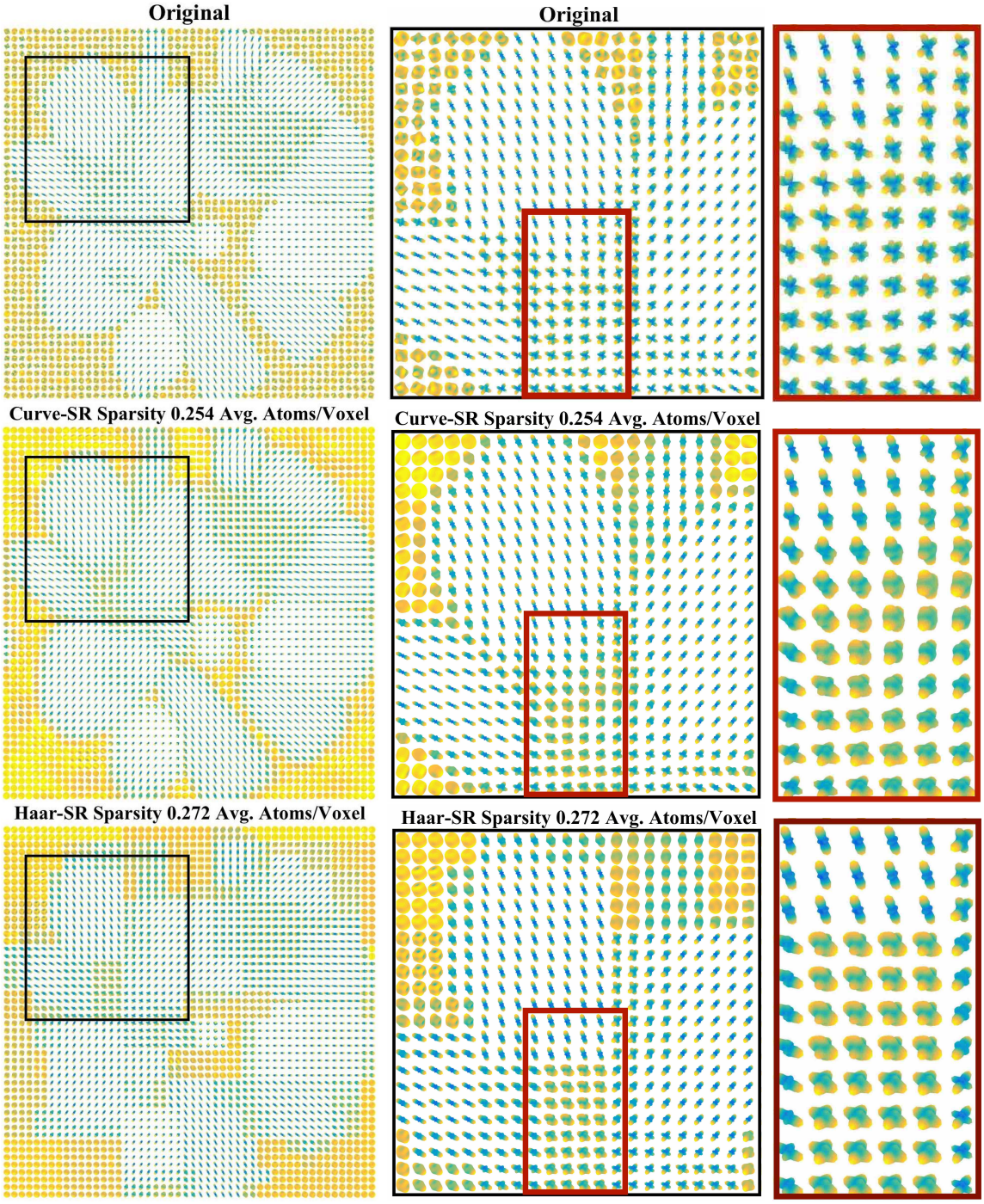


Figure 3.11: Results of the proposed spatial-angular sparse coding using Kron-FISTA for Haar-SR and Curve-SR using an average of ~ 0.25 atoms/voxel compared to original signal. Curve-SR outperforms Haar-SR in this regime due to its additional directionality. We can see a drastically better reconstruction compared to the state-of-the-art at the same sparsity level in the top left of Figure 3.2. This clearly shows that we can achieve accurate reconstruction with less than 1 atom/voxel.

reconstruction with a preservation of smaller detailed fiber tract regions. In contrast, the state-of-the-art I-SR is unable to model intricate fiber regions and is forced to set most voxels to zero atoms. All in all, we can see that using our proposed method, we can achieve much higher sparsity levels than the state-of-the-art, and accurate reconstructions using less than 1 atom/voxel. In terms of efficiency, Kron-FISTA was completed on the real HARDI data of size $V = 60 \times 60 \times 30$, $G = 127$ in 1.5 hours for our sparsity level of interest using the fast 3D wavelet transform implemented in MATLAB.

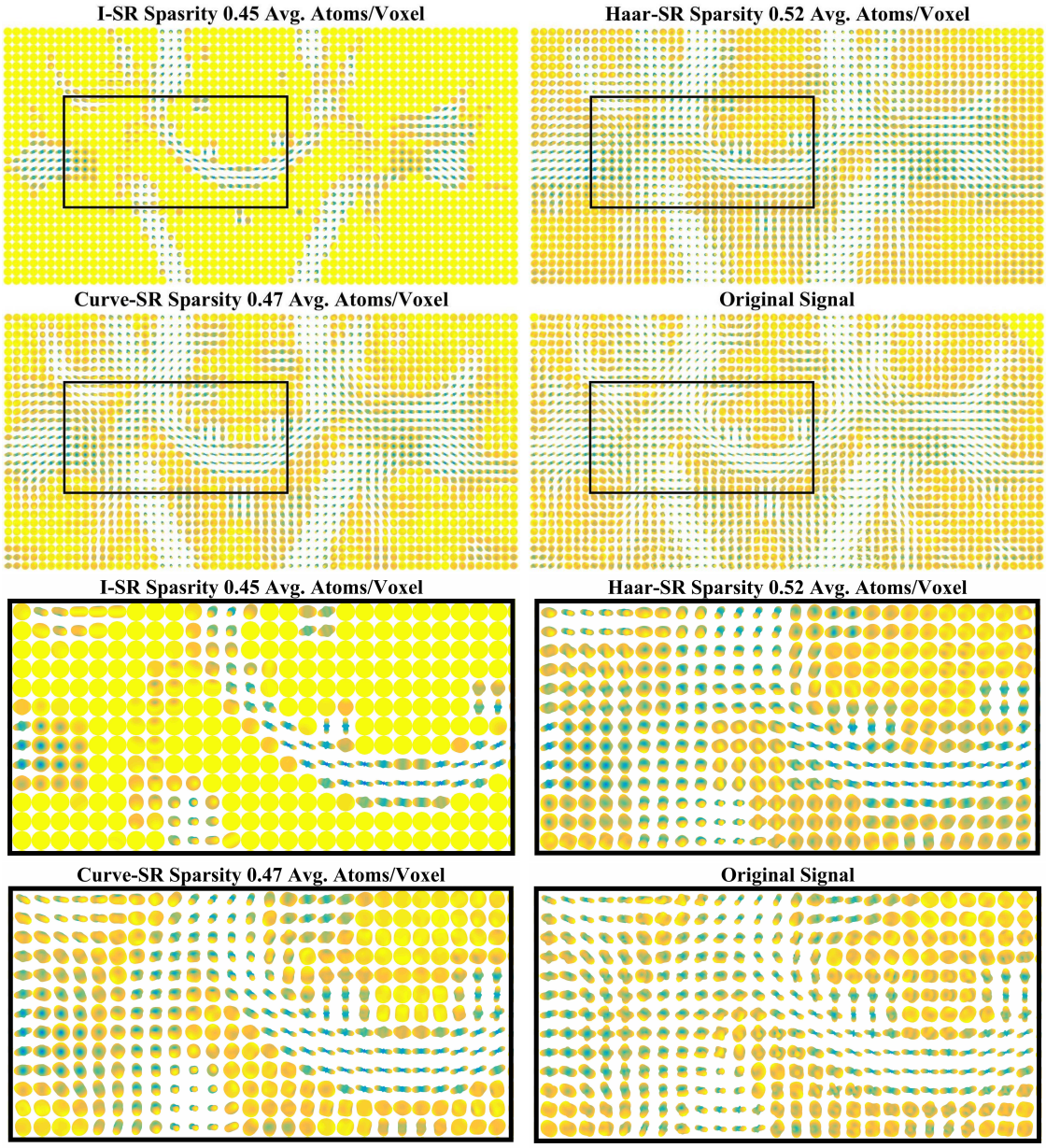


Figure 3.13: Results of proposed spatial-angular sparse coding on real HARDI brain data using Kron-FISTA for I-SR, Haar-SR and Curve-SR at very high sparsity level of ~ 0.5 avg. atoms/voxel compared to original signal. Curve-SR outperforms Haar-SR in this high sparsity range due to its directionality. The state-of-the-art I-SR is unable to compete at this sparsity level.

3.6 Conclusion

In this work, we have demonstrated that by using a joint spatial-angular dictionary, we can obtain accurate HARDI reconstruction with spatial-angular sparsity levels of less than 1 atom per voxel, surpassing the limitations of state-of-the-art angular representations. This provides a new general reconstruction framework to achieve sparser dMRI representations than previously possible with optimal choices of spatial and angular dictionaries. In particular, we have shown promising sparsity results for HARDI from the combination of curvelet (spatial) and spherical ridgelet (angular) dictionaries, but other spatial and angular dictionaries may be chosen for other dMRI protocols like DSI or MS-HARDI.

Furthermore, to efficiently solve this large-scale global sparse coding problem, we have proposed three novel extensions of popular sparse coding algorithms for the Kronecker dictionary setting. All strategies improve upon previously proposed algorithms by explicitly exploiting the separability of the dictionary and each may be beneficial depending on the problem regime and size of data. For our large-scale HARDI data, Kron-FISTA was the leader in speed. Future directions can be to investigate other efficient active set methods such as the recent ORacle Guided Elastic Net (ORGEN) [137].

In addition to sparse coding, our spatial-angular representation may have novel applications in other areas of dMRI processing such as feature extraction, global ODF non-negativity, fiber tract segmentation, and tractography. In Chapter 5, we will

show results of sparse coding on HARDI denoising. Our main application for spatial-angular sparse coding is the promising improvements of acquisition acceleration of dMRI through compressed sensing. In the next chapter, we aim to reduce signal measurements jointly in k- and q-space below the state of the art by naturally incorporating our joint spatial-angular sparse coding within a unified (k, q) compressed sensing framework.

Chapter 4

(k, q)-Compressed Sensing with Spatial-Angular Sparsity

4.1 Introduction

Compressed sensing (CS) [59] has been regularly employed in the literature to accelerate the acquisition of real-world signals and medical images. The main ingredients of the CS framework are an appropriately chosen sampling scheme and an underlying "sparse" representation of the data. In the previous chapter we presented a new spatial-angular representation of dMRI that provides a much sparser reconstruction than the state-of-the-art. In this chapter, we will compare the proposed representation with the state-of-the-art within the CS framework to evaluate the amount of subsampling that can be achieved by each. The key idea is that, with

a well chosen sampling scheme, in general, the sparser the representation, the fewer the samples may be needed to reconstruct the full signal with high accuracy.

CS has been classically applied to MRI [60] by subsampling in the native k-space (k-CS) while applying sparsifying transforms in the spatial image domain like wavelets and total-variation (TV). For dMRI, diffusion signals are measured along different angular gradient directions in q-space for every point in k-space. Thus, to reduce the number of diffusion measurements, many methods [61] have exploited sparse representations in the angular domain by applying CS in q-space (q-CS). To further accelerate dMRI, more recent methods [62, 92, 95, 97] combine aspects of k-CS and q-CS by subsampling jointly in (k, q)-space ((k, q)-CS). However, these methods impose sparsity on the spatial and angular domains *separately*, which can lead to a less efficient representation of dMRI data and may limit the reduction of signal measurements that can be achieved in (k, q)-CS.

In this chapter, we present a new (k, q)-CS framework that subsamples jointly in (k, q)-space while imposing sparsity in the *joint* spatial-angular domain. Building upon the sparse coding findings in Chapter 3 which show increased levels of dMRI sparsity using joint spatial-angular sparse coding, our proposed (k, q)-CS has the potential to further accelerate dMRI than prior methods by exploiting this underlying sparse representation. Our main objective of this chapter is to evaluate the advantages of imposing sparsity in the joint spatial-angular domain versus previous formulations that involve separate spatial and angular sparsity terms. For this reason, our focus

will not yet be the optimization of sparsifying dictionaries or sensing schemes to push the limits of subsampling but first to compare the gains of our proposed model with respect to the state-of-the-art formulations for a fixed setting.

4.2 State-of-the-Art in Compressed Sensing

In this section, we will review the state of the art in compressed sensing (CS), building from the general setting to the specific applications of MRI and dMRI. While, we have reviewed the main ideas of compressed sensing in the background Section 2.2.2, we will heretofor introduce the concepts of *synthesis* and *analysis* models used frequently in the applications MRI and dMRI.

4.2.1 CS for General Signals

In the general setting, a full signal s is reconstructed from undersampled and noisy measurements $\hat{s}$ obtained through an undersampling (or sensing) matrix $\mathcal{U}$ by solving an L_1 minimization program of the form:

$$\min_{s,c} \frac{1}{2}||\mathcal{U}s - \hat{s}||_2^2 + \lambda||c||_1, \tag{4.1}$$

subject to the constraint that **either** $s = \Phi c$ with Φ being a sparsifying dictionary and c the coefficients (*synthesis*) **or** $c = \Phi^\top s$ where $\Phi^\top$ is an analysis operator applied to the signal (*analysis*). Both formulations involve a sparsity prior $\|c\|_1$ in the transform domain of the signal space that is controlled by the balance parameter $\lambda \geq 0$. Note however that in the typical scenario in which Φ is an overcomplete dictionary, *synthesis* and *analysis* CS are not equivalent models (cf. [138] for a thorough discussion). In the *synthesis* case, the optimization is done on the coefficient vector c from which the signal s can be synthesized while in the *analysis* case s is found directly.

4.2.2 k-CS for MRI

One of the first applications of CS has been the acceleration of MRI acquisition [60]. Measurements are made in the frequency domain (called k-space) and the reconstruction is done in the image domain. If we denote by $\hat{s}_k$ the subsampled measurements in k-space and by s_x the fully reconstructed image, the CS problem (4.1) for MRI becomes:

$$\min_{s_x, b} \frac{1}{2}||\mathcal{U}_k \mathcal{F} s_x - \hat{s}_k||_2^2 + \lambda||b||_1, \tag{4.2}$$

subject to the constraint that **either** $s_x = \Psi b$ (*synthesis*) **or** $b = \Psi^\top s_x$ (*analysis*), where $\mathcal{F}$ is the Fourier Transform, $\mathcal{U}_k \in \mathbb{R}^{K \times V}$ is the undersampling k-space matrix, K is the number of samples and V is the total number of voxels with $K \leq V$. Here $\Psi \in \mathbb{R}^{V \times N_\Psi}$ is typically a dictionary of N_Ψ atoms defined on the image domain

(e.g. Wavelets or Curvelets) and $\Psi^\top$ is a sparsifying transform either associated to those dictionaries or to other operators such as the gradient in the case of total variation (TV) regularization. This last choice in the *analysis* formulation ($b = \Psi^\top s_x$) is a common model for sparse MRI reconstruction [60]:

$$\min_{s_x} \frac{1}{2}||\mathcal{U}_k \mathcal{F} s_x - \hat{s}_k||_2^2 + \lambda||\Psi^\top s_x||_1. \tag{4.3}$$

4.2.3 q-CS for dMRI

The structure of dMRI is significantly more complex than that of traditional MRI, whereby for each k-space measurement, a set of G (angular) diffusion measurements are acquired in the analogous q-space. Diffusion signals are traditionally viewed voxel-wise in the image domain (after k-space reconstruction) as a matrix $S_{x,q} = [s_1, \ldots, s_V]^\top \in \mathbb{R}^{V \times G}$, where $s_v \in \mathbb{R}^G$ is the diffusion signal in voxel v. q-CS has been used extensively in the literature [61], each new treatment testing a new sparsifying angular dictionary or sampling scheme. Traditionally formulated as in (4.1) for each voxel v, q-CS is more frequently solved for all voxels simultaneously as:

$$\min_{S_{x,q}, A} \frac{1}{2}||S_{x,q}\mathcal{U}_q^\top - \hat{S}_{x,q}||_F^2 + \lambda||A||_1, \tag{4.4}$$

subject to the constraint that **either** $S_{x,q} = A\Gamma^\top$ (*synthesis*) **or** $A = S_{x,q}\Gamma$ (*analysis*), where $\hat{S}_{x,q} = [\hat{s}_1, \ldots, \hat{s}_V]^\top \in \mathbb{R}^{V \times Q}$ are the measured q-space signals $\hat{s}_v \in \mathbb{R}^Q$ at

each voxel v, $\mathcal{U}_q \in \mathbb{R}^{Q \times G}$ is an undersampling matrix in q-space with $Q \leq G$, and $A = [a_1, \ldots, a_V]^\top \in \mathbb{R}^{V \times N_\Gamma}$ is the matrix of angular coefficients for an angular q-space dictionary $\Gamma \in \mathbb{R}^{G \times N_\Gamma}$ with N_Γ atoms.

Prior work [61] has explored the construction of many sparsifying dictionaries Γ related to estimating orientation distribution functions and so the constraint $S_{x,q} = A\Gamma^\top$ is most commonly used, resulting in the *synthesis* formulation:

$$\min_A \frac{1}{2}||A\Gamma^\top \mathcal{U}_q^\top - \hat{S}_{x,q}||_F^2 + \lambda||A||_1. \tag{4.5}$$

4.2.4 (k, q)-CS for dMRI

A logical advancement to further accelerate dMRI is to additionally subsample in k-space. State-of-the-art methods like [62, 92, 95, 97] have been applied to many dMRI protocols testing various combinations of dictionaries and sensing schemes. Interestingly, all of them can be formulated as particular cases of the following problem, which combine k-CS (4.2) and q-CS (4.4):

$$\min_{S_{x,q}, A, B} \frac{1}{2}||\mathcal{U}_{k,q}(\mathcal{F}S_{x,q}) - \hat{S}_{k,q}||_F^2 + \lambda_1||A||_1 + \lambda_2||B||_1 \tag{4.6}$$

subject to the constraints $S_{x,q} = A\Gamma^\top$ (*synthesis* as in (4.5)) and $B = \Psi^\top S_{x,q}$ (*analysis* as in (4.3)). The sensing scheme $\mathcal{U}_{k,q}$ is now a joint (k, q) subsampling operator (cf. Fig. 4.1 and Sec. 4.6.1 for a discussion) and $\hat{S}_{k,q} \in \mathbb{R}^{K \times Q}$ are the subsampled

measurements in (k, q)-space.

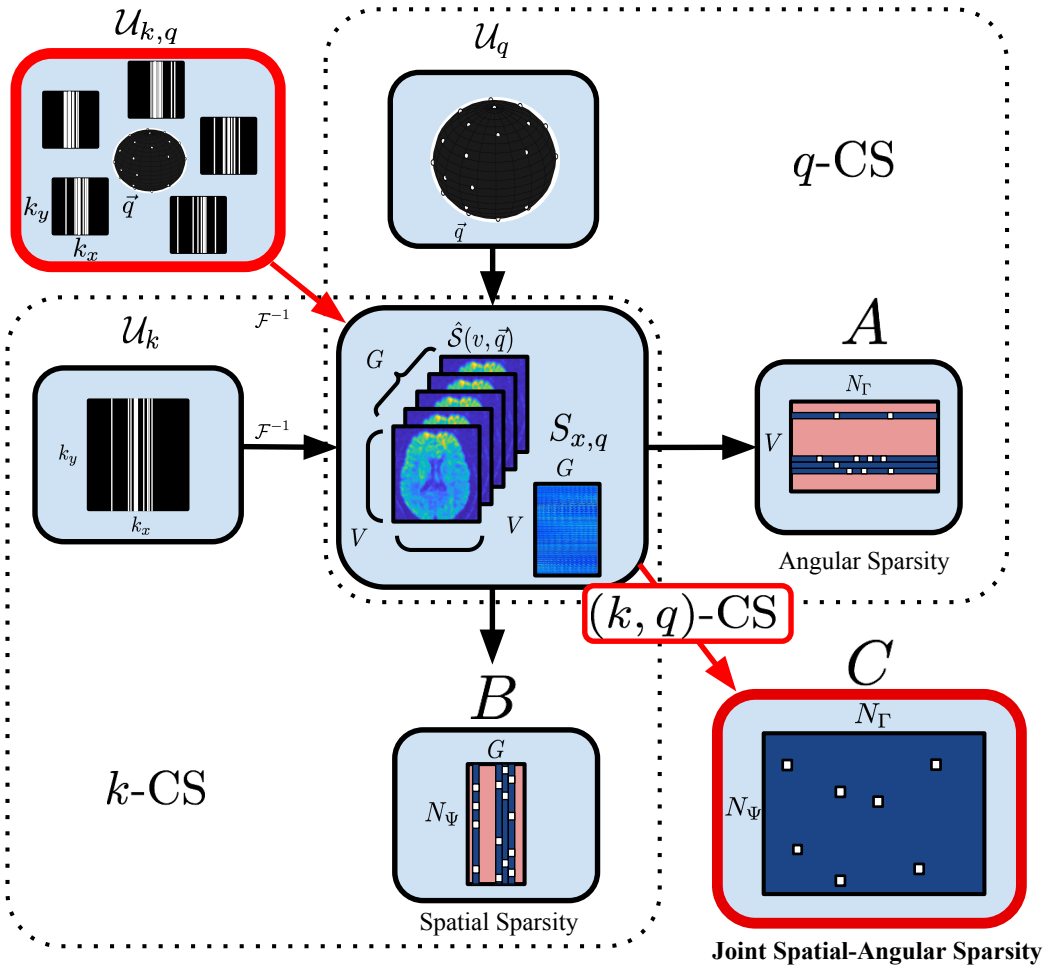


Figure 4.1: Diagram of k-CS, q-CS, and (k, q)-CS with domains of sensing (top left) and sparsity (bottom right). State-of-the-art methods subsample jointly in (k, q)-space with $\mathcal{U}_{k,q}$ but then *add* separate spatial, B (bottom), and angular, A (right), sparsity priors that combine k- and q-CS. Instead, we propose to enforce sparsity in the joint spatial-angular domain, C (bottom-right), resulting in a natural unified framework for (k, q)-CS that allows a reduced number of samples via increased levels of joint sparsity.

As a critical point of distinction, in (4.6) the sparsity prior is imposed on two separate domains: the angular dictionary coefficients $A \in \mathbb{R}^{V \times N_\Gamma}$ at each voxel and the spatial transform coefficients $B \in \mathbb{R}^{G \times N_\Psi}$ for each gradient direction. The sparsity

in these domains, if measured by the L_0 seminorm, is inherently limited by the size of the dMRI data (V,G) since, for non-zero q-space signals at all voxels $||A||_0 \geq V$, and for non-zero k-space images for each gradient direction $||B||_0 \geq G$, resulting in a total spatial plus angular sparsity of $||A||_0 + ||B||_0 \geq V + G$. This sparsity limitation in the model of (4.6) may eventually impact the possible reduction in sampling rate for (k, q)-CS as we will show empirically in our experiments in Section 4.6.

4.3 Proposed (k, q)-CS for dMRI with Joint Spatial-Angular Sparsity

In this section, we propose a new (k, q)-CS model for dMRI involving a single joint spatial-angular sparsity prior (derived in Chapter 3) instead of separate spatial and angular sparsity terms as in (4.6). We consider the full vectorized global signal $s_{x,q} \in \mathbb{R}^{VG}$ as the stacking of each s_x for every q-space point, and a measured subsampled signal in (k, q)-space $\hat{s}_{k,q} \in \mathbb{R}^{KQ}$, such that $\hat{s}_{k,q} = \mathcal{U}_{k,q}(\mathcal{F}s_{x,q})$, where the Fourier transform $\mathcal{F}$ is applied to each spatial component and $\mathcal{U}_{k,q} \in \mathbb{R}^{KQ \times VG}$ is the (k, q) sensing matrix. Then we can write the global (k, q)-CS in vector form, analogous to the general setting in (4.1):

$$\min_{s_{x,q},c} \frac{1}{2}||\mathcal{U}_{k,q}(\mathcal{F}s_{x,q}) - \hat{s}_{k,q}||_2^2 + \lambda||c||_1, \tag{4.7}$$

subject to the constraint that **either** $s_{x,q} = \Phi c$ (*synthesis*) **or** $c = \Phi^\top s_{x,q}$ (*analysis*), where $\Phi \in \mathbb{R}^{VG \times N_\Phi}$ is a global sparsifying dictionary. Notice that (4.7) has a direct statistical interpretation as a reconstruction under a sparsity prior with respect to the dictionary Φ. However, numerically solving such an optimization problem is largely intractable due to the size of dMRI data ($|s_{x,q}| = VG \approx 100^4$) and the resulting huge size of Φ.

As was proposed for spatial-angular sparse coding in Chapter 3, to overcome this computational difficulty, we choose Φ to be separable over the spatial and angular domains resulting in the Kronecker dictionary $\Phi = \Gamma \otimes \Psi$, where Ψ and Γ are the spatial and angular dictionaries, respectively. Then, we can rewrite (4.7) in an equivalent matrix form as:

$$\min_{S_{x,q}, C} \frac{1}{2} ||\mathcal{U}_{k,q}(\mathcal{F} S_{x,q}) - \hat{S}_{k,q}||_F^2 + \lambda ||C||_1, \tag{4.8}$$

subject to the constraint that **either** $S_{x,q} = \Psi C \Gamma^\top$ (*synthesis*) **or** $C = \Psi^\top S_{x,q} \Gamma$ (*analysis*). In fact, substituting also the constraints from k-CS (4.2) and q-CS (4.4), a separable spatial-angular dictionary allows us to have two additional constraint options: (1) $S_{x,q} = A\Gamma^\top$ **and** $C = \Psi^\top A$ (*analysis-synthesis*) **or** (2) $S_{x,q} = \Psi B$ **and** $C = B\Gamma$ (*synthesis-analysis*).

Notice that, in contrast to the state-of-the-art formulation in (4.6), our formulation only involves one penalty term that imposes sparsity on the *joint* spatial-angular coefficient domain $C \in \mathbb{R}^{N_\Psi \times N_\Gamma}$ of the global dictionary $\Gamma \otimes \Psi$ (cf. Fig. 4.1). The sparsity of this domain is *a priori* not limited by the size of the data and so this joint

	Sensing	Signal		Coefficients		
Variable	$\mathcal{U}_{k,q}$	$\hat{S}_{k,q}$	$S_{x,q}$	A	B	C
Dimensions	$V \times G \rightarrow K \times Q$	$K \times Q$	$V \times G$	$V \times N_\Gamma$	$G \times N_\Psi$	$N_\Psi \times N_\Gamma$

Table 4.1: Compressed sensing variable dimensions, where G (≈ 100) is the full number of gradient directions in q-space, $Q \ll G$ is the number of measured samples in q-space, V ($\approx 100^3$) is the number of voxels in the volume, $K \ll V$ is the number of measured samples in k-space, N_Γ ($\gtrsim 100$) is the number of atoms of the angular dictionary Γ, and N_Ψ ($\gtrsim 100^3$) is the number of atoms of the spatial dictionary Ψ. A are the angular coefficients per voxel, B are the spatial coefficients per gradient direction, and C are the spatial-angular coefficients.

model can lead to sparser representations of typical dMRI signals than summing separate spatial and angular terms. In the next section, we present an algorithm to efficiently solve the proposed (k, q)-CS formulation.

In contrast with the usual formulation (4.6) which essentially constrains sparsity through separate spatial and angular terms, the approach we will follow for dMRI Compressed Sensing is in direct continuity with the joint sparsity model presented in Chapter 3, in which sparsity of the reconstructed signal is imposed on a joint spatial-angular domain we call C. Fig. 4.1 depicts a full schematic summary of the domains of sampling in k-CS, q-CS, and the joint (k, q)-CS, and the associated sparsity priors in the spatial domain (B), angular domain (A) and the joint spatial-angular (C) domain.

As motivated by Fig. 4.1, while A is row-sparse ($||A||_0 \geq V$), and B is column sparse ($||B||_0 \geq G$), C has no *a priori* structured sparsity ($||C||_0 \geq 1$), meaning that our formulation has the potential to achieve greater sparsity levels and therefore higher subsampling rates within (k, q)-CS than the state of the art. We can actually

give a slightly more precise justification of our proposed heuristic based on the existing theory of CS. Although there are multiple results associated to different formulations of CS or the structure of the involved dictionaries, we will simplify our discussion by considering the case of redundant dictionaries (which is the typical situation in our applications) and reconstruction through the analysis model.

4.4 Heuristic Comparison of the Separate and Joint Sparsity Priors

In this section, we wish to motivate some heuristic intuition for comparing the state-of-the-art compressed sensing with separate spatial and angular sparsity penalties and the proposed joint spatial-angular sparsity penalty. We adapt the ideas of D-RIP presented in the background Section 2.2.2.3 to better evaluate the theoretical differences between the two models.

Consider again two given dictionaries $\Psi \in \mathbb{R}^{V \times N_\Psi}$ and $\Gamma \in \mathbb{R}^{G \times N_\Gamma}$ (that we assume to be tight frames) in the spatial and angular domains respectively, as well as a matrix signal $S \in \mathbb{R}^{V \times G}$ to reconstruct from its measurements $\mathcal{U}(S) = \hat{S} \in \mathbb{R}^{K \times Q}$ with $K \ll V$ and $Q \ll G$ (U includes the Fourier transform of the spatial part to simplify notations). The analysis basis pursuit reconstruction problems in both the

separate and joint formulations may be rewritten similarly to (2.53) as

$$\hat{s} = \arg\min_{\hat{s}} \|D^* \hat{s}\|_1 \quad \text{subject to} \quad \|\hat{s} - y\|_2 \leq \epsilon$$

where $s \in \mathbb{R}^{VG}$ and $\hat{s} \in \mathbb{R}^{KQ}$ are the vectorized versions of S and $\hat{S}$ and the dictionary D is given respectively by $D = \Gamma \otimes \Psi \in \mathbb{R}^{VG \times N_\Gamma N_\Psi}$ in the joint sparsity case or as $D = [\Gamma \otimes \mathrm{I}_{V \times V} \quad \mathrm{I}_{G \times G} \otimes \Psi] \in \mathbb{R}^{VG \times (N_\Gamma V + G N_\Psi)}$ in the separate case. Indeed, it is easy to check that $\|D^* \hat{s}\|_1 = \|(\Gamma \otimes \Psi)^* \hat{s}\|_1 = \|\Psi^\top \hat{S} \Gamma\|_1$ in the joint case and $\|D^* \hat{s}\|_1 = \|\Psi^\top \hat{S}\|_1 + \|\hat{S}\Gamma\|_1$ in the separate case. Note that these two decomposition operators take values in spaces of different dimensions: $N_\Gamma N_\Psi$ and $N_\Gamma V + N_\Psi G$ respectively.

Now, from the result of Theorem 6 in the background Chapter 2.2.2.3, in both formulations, assuming the sensing matrix $\mathcal{U}$ satisfies the adequate D-RIP property with $\delta_{2J} < 0.08$ for some J, we have that the reconstruction error is bounded by $\|\tilde{s} - s\|_2 \leq C_0 \epsilon + C_1 \frac{\|D^* s - (D^* s)_J\|_1}{\sqrt{J}}$ (with possibly different constants). We argue however that generically, with the right choice of dictionaries, the sparsity properties of signals in the joint domain can be much better than in the separate domain, in other words that the term $\frac{\|D^* s - (D^* s)_J\|_1}{\sqrt{J}}$ may be significantly smaller in the former case. This is already in part supported by our previous experiments on sparse coding. But more fundamentally, it is a consequence of the fact that using separate sparsity terms can only exploit redundancies of the q-space signal at each voxel separately and of the x-

space signal corresponding to each diffusion gradient separately again. Let's consider the simplest example of a constant signal s over all voxels and gradients, and assume that the signal is sparsely decomposed with $Q \ll G$ atoms in Γ at each voxel and $K \ll V$ atoms in Ψ for each gradient direction. With the separate dictionary, it results in D^*s having $QV + KG$ non-zero coefficients in total and thus, to get exact recovery from Theorem 6, one would *a priori* need $J \geq QV + KG$.

For the joint dictionary $D = \Gamma \otimes \Psi$ on the other hand, the transform of s by the angular dictionary Γ being constant over all voxels with Q non-zero coefficients, provided constant signals are K-sparse in the spatial dictionary Ψ, D^*s will essentially have KQ non-zero coefficients which is significantly less than the previous bound. Figure 4.2 further illustrates this fact by showing the rates of decrease with respect to the percentage of nonzero coefficients J/N_D. This simulation was done for a simple synthetic HARDI dataset with block regions of constant signal with crossing fibers. It is again evident that the rate of decrease of $\|D^*s - (D^*s)_J\|_1$ with D being equal to a joint dictionary is much faster than using sum of sparsity priors.

Note however that the reconstruction bound of Theorem 6 only holds under the assumption that the undersampling matrix $\mathcal{U}$ is adapted to the dictionary through the D-RIP constraint $\delta_{2J} < 0.08$. This condition thus depends on the dictionary itself and since computing numerically the constants δ_J is of combinatorial complexity, it is can be a particularly difficult condition to verify in practical situations. Yet, for certain classes of random matrices, it can be shown to be satisfied with overwhelming

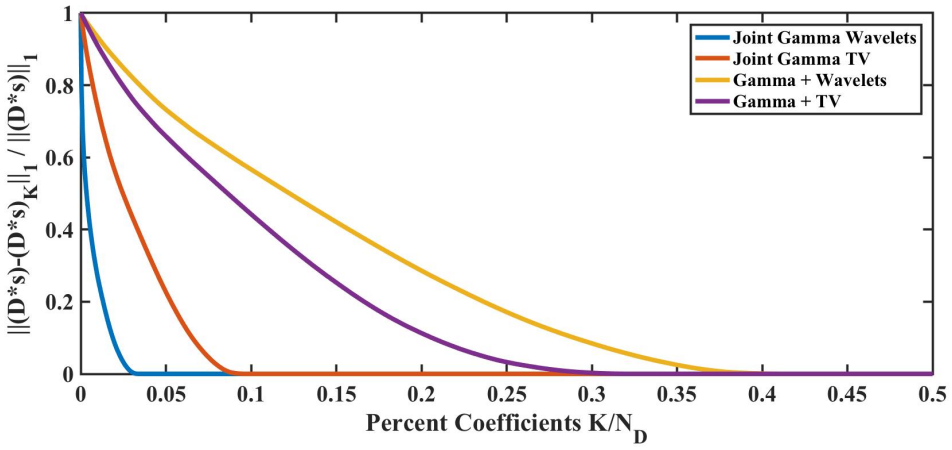


Figure 4.2: Illustration of the D-RIP condition on simple synthetic HARDI data. Joint dictionaries exhibit a higher rate of decrease than separate dictionaries meaning that (with appropriate sensing) the number of measurements needed for accurate signal recovery is expected to be less for joint dictionaries.

probability, leading to universal reconstruction results in those cases. For example, as stated in [71], for a dictionary D of size $VG \times N_D$, a random Gaussian or Bernoulli matrix $\mathcal{U} \in \mathbb{R}^{M \times VG}$ will satisfy the D-RIP condition with overwhelming probability as long as the number of measurements M is such that $M \gtrsim J \log(N_D/J)$. Then Theorem 6 essentially shows that the number of measurements needed for accurate recovery of s is directly related to the speed of decrease of $\|D^*s - (D^*s)_J\|_1$ with respect to J.

Similar conclusions hold for many other random sampling matrices thanks to the result of [139] where it is shown that matrices satisfying the standard RIP property with randomized column signs also satisfy D-RIP. In particular, random subsampled Fourier matrix with randomized signs verifies D-RIP with overwhelming probability for $M \gtrsim J \log^4(VG)$. For all these cases, the theory of CS combined with the

increased sparsity of the joint model suggest that fewer measurements should be eventually needed for approximate recovery of the original signal, which we shall confirm experimentally in Section 4.6.

4.5 Efficient Algorithm to Solve (k, q)-CS

Prior work such as [62,92,95] each solve (4.6) using the Split-Bregman/Alternating Direction Method of Multipliers (ADMM) algorithm and divide the reconstruction per voxel. In this section, we propose an efficient algorithm to solve (k, q)-CS globally for large-scale dMRI data. The proposed algorithm can easily be applied to both the prior formulation (4.6) and our proposed formulation (4.8).

We begin by taking care of the constraints to eliminate variables and simplify the problems. For (4.6), we substitute the prior methods' selected constraints $S_{x,q} = A\Gamma^\top$ and $B = \Psi^\top S_{x,q} = \Psi^\top A\Gamma^\top$ to get:

$$\min_A \frac{1}{2}||\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top) - \hat{S}_{k,q}||_F^2 + \lambda_1||A||_1 + \lambda_2||\Psi^\top A\Gamma^\top||_1. \qquad \text{(Prior)}$$

In order to directly compare our proposed framework (4.8) with (Prior) in terms of variable A, we substitute $S_{x,q} = A\Gamma^\top$ and $C = \Psi^\top A$ (*analysis-synthesis*) to get:

$$\min_A \frac{1}{2}||\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top) - \hat{S}_{k,q}||_F^2 + \lambda||\Psi^\top A||_1. \qquad \text{(SAAS)}$$

We call this formulation **Spatial-Angular Analysis-Synthesis (SAAS)** due to the resulting *analysis* formulation for the spatial domain and *synthesis* formulation for the angular domain. While these substitutions mask the domains of sparsity by using a common variable $A \in \mathbb{R}^{V \times N_\Gamma}$, note that the proposed formulation (SAAS) still imposes sparsity on the joint spatial-angular domain in contrast to the separate spatial and angular sparsity terms of (Prior).

The Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [103] has been well studied for solving L_1 *synthesis* minimization problems such as (4.1), where the proximal operator of $||c||_1$ is the well-known shrinkage function. However, in the *analysis* setting, the proximal operator of a linearly transformed variable such as $||\Psi^\top A||_1$ in (SAAS) and $||\Psi^\top A \Gamma^\top||_1$ in (Prior) is not directly computable. There are multiple ways to overcome this. In particular, [140] proposes a method that applies FISTA to a relaxed *smooth* problem, coined Smooth FISTA (SFISTA). In what follows, we adapt SFISTA to the separable Kronecker matrix setting in order to solve (SAAS) and (Prior).

First, (SAAS) is reformulated by introducing the auxiliary linear constraint $Z = \Psi^\top A$ and the unconstrained relaxed optimization becomes:

$$\min_{A,Z} \frac{1}{2}||\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top) - \hat{S}_{k,q}||_F^2 + \lambda||Z||_1 + \frac{\rho}{2}||Z - \Psi^\top A||_F^2. \tag{4.9}$$

Let $f(A) \equiv ||\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top) - \hat{S}_{k,q}||_F^2$. Since f does not depend on Z, we can pass the

minimization with respect to Z to the last two terms. Define $g_\mu(X) \equiv \min_Z ||Z||_1 + \frac{1}{2\mu}||Z - X||_F^2$. Then (4.9) is equivalent to

$$\min_A f(A) + \lambda g_{\frac{\lambda}{\rho}}(\Psi^\top A). \tag{4.10}$$

Here g_μ is the Moreau envelope of the L_1 norm which can be shown to equal the Huber function given by $\mathcal{H}_\mu(x) = \frac{1}{2\mu^2}x^2$ if $|x| < \mu$ and $|x| - \frac{\mu}{2}$ otherwise. We can now apply FISTA to the smooth (4.10) by taking an accelerated gradient descent using

$$\nabla f(A) = \mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top))\Gamma - \mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\hat{S}_{k,q})\Gamma \tag{4.11}$$

$$\nabla g_{\frac{\lambda}{\rho}}(\Psi^\top A) = \frac{\rho}{\lambda}\Psi(\Psi^\top A - \text{shrink}_{\frac{\lambda}{\rho}}(\Psi^\top A)) \tag{4.12}$$

where $\mathcal{U}_{k,q}^*$ is the operator that restores the subsampled signal to full size by filling in unsampled indices of the full signal with zeros (cf. Sec. 4.6.1 for a discussion).

As discussed for our derivation of Kron-FISTA in Section 3.4.5, FISTA is guaranteed to converge using a step-size L that is greater than or equal to a Lipschitz constant of the objective f. For the case of SFISTA, the Lipschitz constants of f and

g are added [140]. Using (4.11), a Lipschitz constant of f is:

$$\begin{aligned}||\nabla f(A_1) - \nabla f(A_2)||_F &= ||\mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\mathcal{U}_{k,q}(\mathcal{F}A_1\Gamma^\top))\Gamma - \mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\mathcal{U}_{k,q}(\mathcal{F}A_2\Gamma^\top))\Gamma||_F \\ &= ||\mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\mathcal{U}_{k,q}(\mathcal{F}(A_1 - A_2)\Gamma^\top))\Gamma||_F \\ &= ||\mathcal{F}^{-1}\mathcal{U}_{k,q}^*(\mathcal{U}_{k,q}(\mathcal{F}(A_1 - A_2)\Gamma^\top))\Gamma||_F.\end{aligned}$$

The proposed Kronecker SFISTA (Kron-SFISTA) is presented in Algorithm 10.

Algorithm 10 Kron-SFISTA for SAAS model (k, q)-CS

Choose: λ, ρ, ϵ.
Initialize: $i = 1, A_0 = Y_1 = \mathbf{0}$, $n_1 = 1, L \geq \lambda_{\max}(\Gamma^\top\Gamma) + \rho\lambda_{\max}(\Psi\Psi^\top)$.
while error $> \epsilon$ **do**
1: $A_i = Y_i - (\nabla f(Y_i) + \lambda\nabla g_{\lambda/\rho}(\Psi^\top Y_i))/L$;
2: $n_{i+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_i^2})$;
3: $Y_{i+1} = A_i + \frac{n_i - 1}{n_{i+1}}(A_i - A_{i-1})$;
4: $i \leftarrow i + 1$;
end while
Return: $\hat{A}$.
Reconstruct: $\hat{S}_{x,q} = \hat{A}\Gamma^\top$.

According to [140], we choose stepsize $L \geq \lambda_{\max}(\Gamma^\top\Gamma) + \rho\lambda_{\max}(\Psi\Psi^\top)$ to guarantee convergence, where $\lambda_{\max}(X)$ is the max eigenvalue of X. The parameter ρ is gradually increased using parameter continuation [140] to ensure convergence. The trade-off parameter λ, dictates the level of sparsity of $\Psi^\top A$. A large value of λ will result in a very sparse representation at the expense of reconstruction accuracy, while a small value of λ may result in over-fitting the sampled data at the expense of reconstruction accuracy of unseen data. Therefore, in our experiments we vary the level of λ and select the value that leads to a minimal reconstruction error. The efficiency of Kron-

SFISTA over the traditional SFISTA can be viewed in the same vein as for Kron-FISTA analyzed in [85].

As an alternative to the frequently used Split-Bregman, Kron-SFISTA can also be easily applied to (Prior) by solving:

$$\min_A f(A) + \lambda_1 ||A||_1 + \lambda_2 g_{\frac{\lambda_2}{\rho_2}}(\Psi^\top A \Gamma^\top). \tag{4.13}$$

Step 1 in Alg. 10 becomes $A_i = \text{shrink}_{\lambda_1/L}(Y_i - (\nabla f(Y_i) + \lambda_2 \nabla g_{\frac{\lambda_2}{\rho_2}}(\Psi^\top Y_i \Gamma^\top))/L)$ with $\nabla g_{\frac{\lambda_2}{\rho_2}}(\Psi^\top A \Gamma^\top) = \frac{\rho_2}{\lambda_2}\Psi(\Psi^\top A \Gamma^\top - \text{shrink}_{\frac{\lambda_2}{\rho_2}}(\Psi^\top A \Gamma^\top))\Gamma$. This provides an efficient global algorithm to solve (k, q)-CS for large-scale dMRI data.

4.6 Experiments on (k, q)-CS

In our experiments, we focus on two main objectives. The first objective, in Section 4.6.2, is to directly compare the reconstruction accuracy of (SAAS) and (Prior) for various rates of subsampling with experimental results on phantom and real HARDI brain data in Section 4.6.2.1 and Section 4.6.2.2, respectively. The second objective, in Section 4.6.3, is to confirm the ability of the proposed method to generalize to a large set of real HARDI brain data with parameters tuned from a single training subject as may be the setting for future experiments with the proposed methodology on newly acquired data.

We postpone optimizing the amount of subsampling to future work and there-

fore explore somewhat classical choices of spatial and angular dictionaries/transforms and sensing schemes previously tested in the literature. We present our choices of dictionaries and sampling schemes first in Section 4.6.1.

4.6.1 Spatial-Angular Transforms and (k, q) Subsampling Schemes

Spatial Transform $\Psi^\top$. For spatial transform $\Psi^\top$, we consider in our experiments two popularly used transforms for k-CS: Haar wavelets and the finite difference (gradient) operator $\nabla = [\partial_x, \partial_y, \partial_z]$. In the case of the gradient transform, we consider the norm given by $||\nabla(X)||_{2,1} = ||\sqrt{|\partial_x X|^2 + |\partial_y X|^2 + |\partial_z X|^2}||_1$, known as isotropic TV (isoTV)[1]. These transforms have been classically used to sparsely represent MRI images.

Angular Dictionary Γ. The choice of angular dictionary Γ depends on the q-space acquisition protocol of the data. For example, Γ must be chosen to model Cartesian sampled q-space signals for DSI, and multi-shell q-space signals with a radial component for multi-shell HARDI. It is important to note our framework is general to any q-space acquisition protocol with an appropriate choice of Γ. In our experiments we use single-shell HARDI data and choose the over-complete spherical

[1]SFISTA must be changed slightly to incorporate the $||\cdot||_{2,1}$ proximal operator $\text{shrink}_\mu^{2,1}(X) = \frac{X}{||X||_{2,\cdot}} \max(||X||_{2,\cdot} - \mu, 0)$ [141], where $||X||_{2,\cdot}$ indicates taking the 2-norm of the columns of X. Its Moreau envelope is $g_\mu^{2,1}(X) \equiv \min_Z ||Z||_{2,1} + \frac{1}{2\mu}||Z - X||_F^2 = \frac{1}{2\mu^2}||X||_{2,\cdot}^2$ if $||X||_{2,\cdot} < \mu$ and $||X||_{2,\cdot} - \frac{\mu}{2}$ otherwise.

ridgelet (SR) dictionary [44], which has been shown to sparsely model HARDI signals. With this comes the spherical wavelet (SW) dictionary for which we can estimate orientation distribution functions (ODFs) from the SR coefficients. With our choice of parameters, this results in $N_\Gamma = 1169$ atoms from which we may choose any subset greater than G for an overcomplete dictionary.

Joint (k, q) Subsampling Scheme $\mathcal{U}_{k,q}$. We experiment with different subsampling schemes in the k and q domains. For subsampling in k-space, the vast literature on k-CS provides many established sensing schemes such as random sampling, radial sampling, and spiral sampling. Many comparisons have been made between different types of sampling in light of physical constraints dictated by the programming of an MRI scanner. For simplicity, we choose a commonly used k-space sampling scheme of constant lines along the k_y direction. The k_x location of the line samples were chosen randomly with respect to a variable density function centered around the zero-frequency location. Next, in q-space, the options for sampling HARDI include semi-uniform subsampling and random sampling on the sphere. We choose the latter, a random subsampling of the points on the sphere for the benefit of CS.

Then, to combine the subsampling in k- and q-space we have two main options. The first is a separable sensing scheme for which the same k-space subsampling is taken for each sampled q-space point. Mathematically, $\mathcal{U}_{k,q} = \mathcal{U}_k \otimes \mathcal{U}_q$ and $\mathcal{U}_{k,q}(\mathcal{F}A\Gamma^\top) = \mathcal{U}_k \mathcal{F} A \Gamma^\top \mathcal{U}_q^\top$ where $\mathcal{U}_k \in \mathbb{R}^{K \times V}$ and $\mathcal{U}_q \in \mathbb{R}^{Q \times G}$. Algorithmically, this makes the computation of $\mathcal{U}_{k,q}^* = \mathcal{U}_{k,q}^\top = \mathcal{U}_k^\top \otimes \mathcal{U}_q^\top$ straightforward. However, separa-

ble sensing strategies may not fully exploit the potentialities of (k, q) subsampling as some experiments in [97] show.

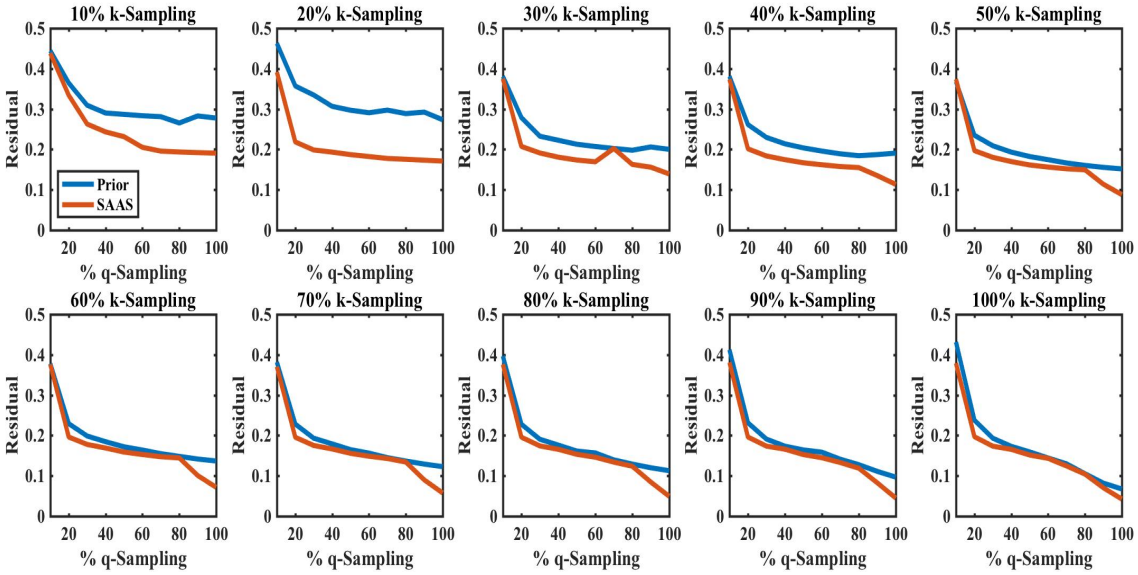


Figure 4.3: Residual error vs. percentage of (k, q) subsampling of the 2D Phantom HARDI data using isoTV and SR for (SAAS) (red) and (Prior) (blue). (SAAS) provides more accurate reconstruction, especially at lower levels of (k, q) subsampling (top left plots).

Alternatively, a non-separable sensing scheme in which a different k-space sampling is used for each sampled q-space point has been proposed [97]. Intuitively, non-separable sensing increases the range of uniquely sampled points and the level of randomness, which are beneficial in CS. In this case, $\mathcal{U}^*_{k,q}$ is the operator that restores the subsampled signal to full size by zeroing out unsampled indices of the full signal. Our implementation of Kron-SFISTA has the benefit of being able to easily handle this non-separable sensing operator, but computationally this is not straight forward in alternative algorithmic formulations such as a Kron-ADMM [85], for example. We compare the reconstruction performances of separable vs. non-separable sensing in

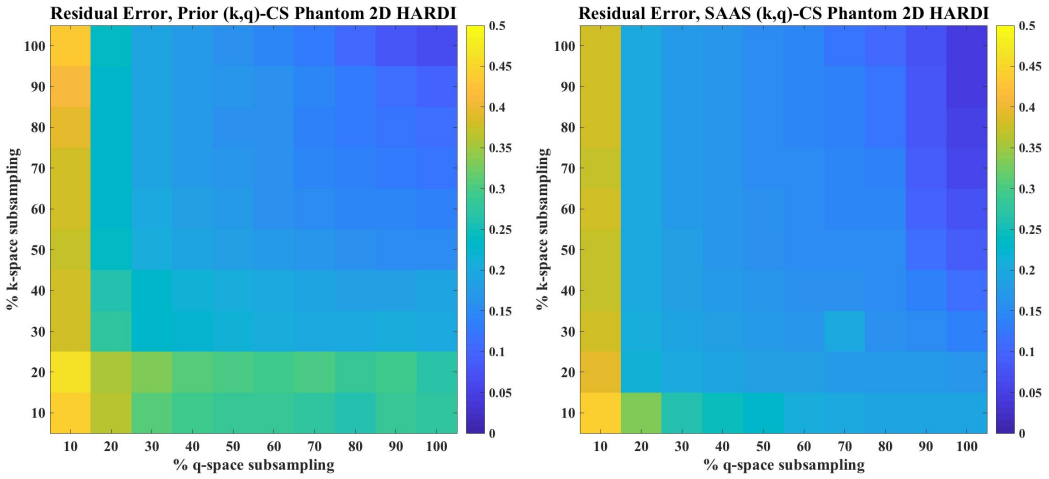


Figure 4.4: Residual error as a function of (k, q) subsampling percentage for the 2D Phantom HARDI data using isoTV and SR. This is another visualization of the data in Figure 4.3. For the Prior method (left), it appears that the amount of error is more symmetrical between subsampling in k- vs. q-space than the error using SAAS (right) which increases more sharply as k-space subsampling is increased. This can be seen by following the change in error along the rows and columns of each plot.

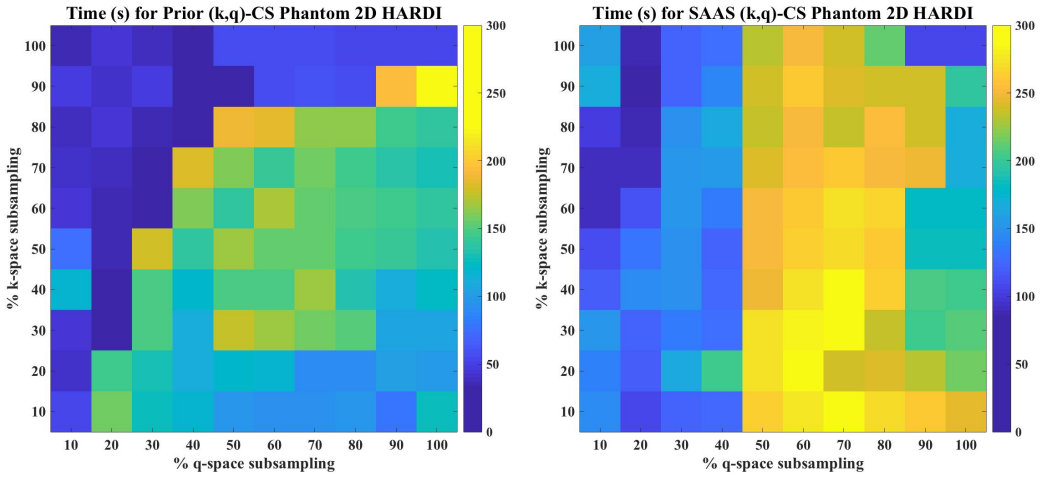


Figure 4.5: Computation time in seconds as a function of (k, q) subsampling percentage for the 2D Phantom HARDI data using isoTV and SR in Figure 4.3. For the Prior method, the computation time is shorter when either k-space has full sampling, or q-space has more undersampling. For the proposed SAAS method, the computation times are more dependant on the q-space sampling alone, shorter when q-space subsampling below 50% and again somewhat shorter at full sampling.

our second set of experiments on a large number of real HARDI subjects.

As a note, the data used in these experiments have been pre-transformed to the spatial domain from the raw k-space data and so we retrospectively transform the data back to k-space using the Fourier transform in order to subsample the data before experiments. Applying our methods directly to raw data acquired in (k, q)-space will be the focus of future work on real data.

4.6.2 Proposed vs. State-of-the-Art (k, q)-CS

4.6.2.1 Phantom HARDI Data

First, we applied our methods on the ISBI 2013 HARDI Reconstruction Challenge Phantom dataset[2], a $V = 50 \times 50 \times 50$ volume with $G = 64$ gradient directions ($b = 3000$ s/mm^2) and SNR = 30, which consists of 20 phantom fibers crossing within an inscribed sphere. We experimented on a middle 2D 50×50 slice of this data. In this experiment, we vary the percentage of subsampling in both k- and q-space, ranging from 10% to 100% of the original phantom HARDI signal in each domain, resulting in a combined total of 1% to 100% of the full signal. Then, we compare our reconstructed signal $\hat{S}_{x,q}$ with the original full signal, $S_{x,q}$, by calculating residual error $||\hat{S}_{x,q} - S_{x,q}||_2^2 / ||S_{x,q}||_2^2$. As a note, this phantom data has been pre-transformed to the spatial domain and so to test k-space subsampling, we retrospectively transform the

[2] http://www.hardi.epfl.ch/static/events/2013_ISBI/testing_data.html

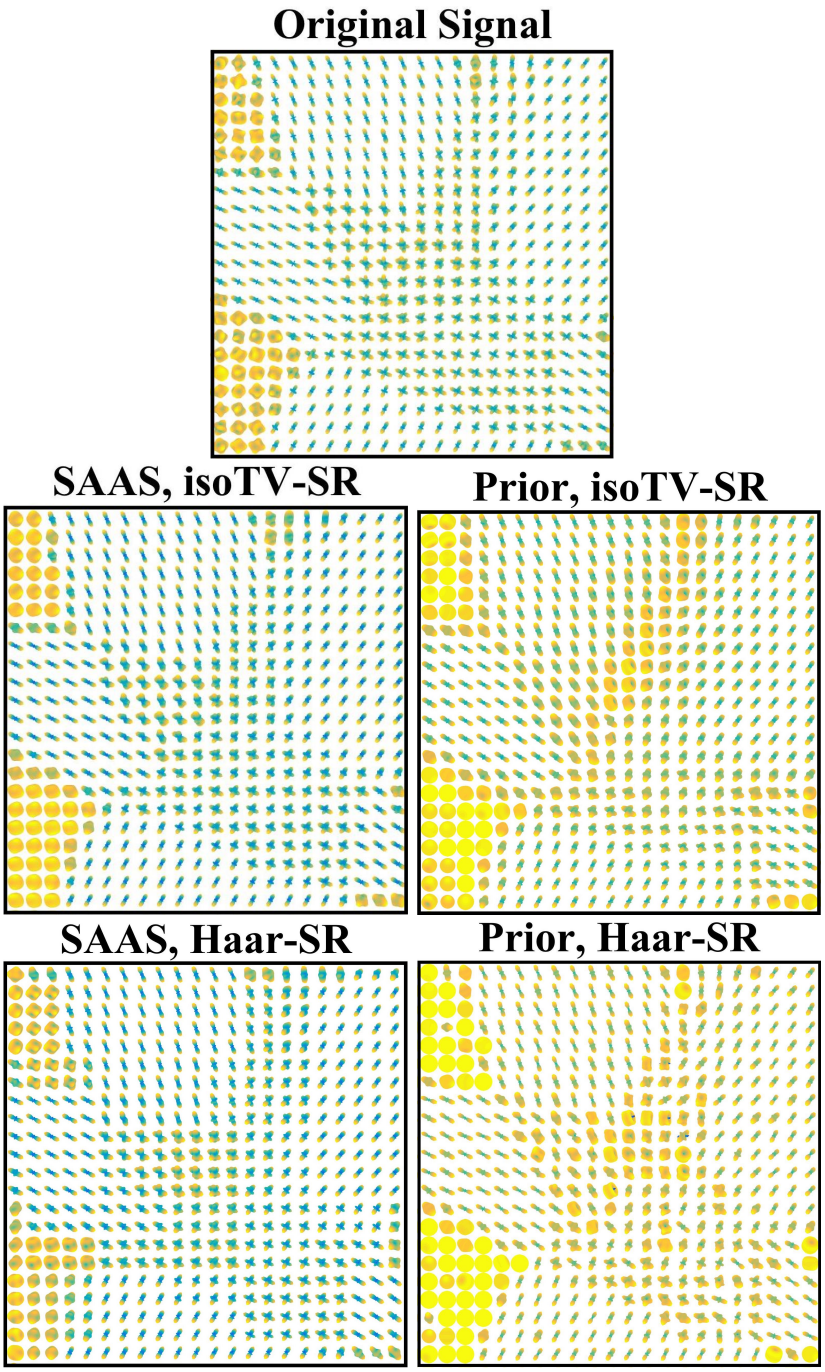


Figure 4.6: Estimation of ODFs from reconstructed phantom signals compared to the original fully sampled signal using the proposed (SAAS) and (Prior). Each is reconstructed from 4% total (k, q) measurements, keeping 20% k-space samples and 20% q-space samples. It is apparent that the prior model is unable to accurately reconstruct crossing fiber signal in the middle of the image. It is also evident that isoTV outperforms Haar.

data back to k-space using the Fourier transform before running different algorithms.

In Fig. 4.3 are the quantitative reconstruction results of our proposed (SAAS) (k, q)-CS compared to (Prior). Each subplot presents a fixed k-space subsampling percentage, while the percentage of q subsampling varies along the x-axis. Kron-SFISTA took $\sim$15-30 min to complete for a sequence of 20 values of λ. We can see improvements of reconstruction accuracy for our proposed method especially in the desired low range of 20% k subsampling and 20% q subsampling, i.e. 500 frequency measurements and 12 gradient directions, keeping a total of 4% of samples (see second plot in first row of Fig. 4.3). The results are visualized as heat maps for comparison in Fig. 4.4. The computation time for each sampling rate are shown in Fig. 4.5

We show the ODFs estimated from the reconstructed phantom signal for this 4% sampling rate in Fig. 4.6 comparing the results of using isoTV versus Haar wavelets. We notice that (Prior) is unable to reconstruct the complex crossing fiber ODFs in the middle region of the image at this low level of sampling. Alternatively (SAAS) provides more accurate reconstructions of the entire dataset with isoTV well outperforming Haar wavelets.

4.6.2.2 Real HARDI Brain Data

We next show (k, q)-CS results on a real HARDI brain dataset with $G = 256$ gradient directions ($b = 1500$ s/mm^2). For visualization we tested on a 2D 50×50 sagittal slice of the corpus callosum region known for two distinct crossing fiber tract

populations in the left-right and anterior-posterior directions. Fig. 4.7 shows the results of our proposed (SAAS) vs. (Prior) first with 20% k-space and 20% q-space (51 gradient directions) subsampling and then decreased to 20% k-space and 10% q-space (25 gradient directions) for a total of 4% and 2% subsampling, respectively. We can see that at 4%, (SAAS) is able reconstruct the crossing ODFs in this region while (Prior) results in isotropic estimations. As we decrease subsampling further to 2%, we notice that (Prior) produces a highly inaccurate reconstruction, setting many voxels to zero (yellow spheres). (SAAS) maintains a recognizable structure but begins to lack accuracy of crossing fibers.

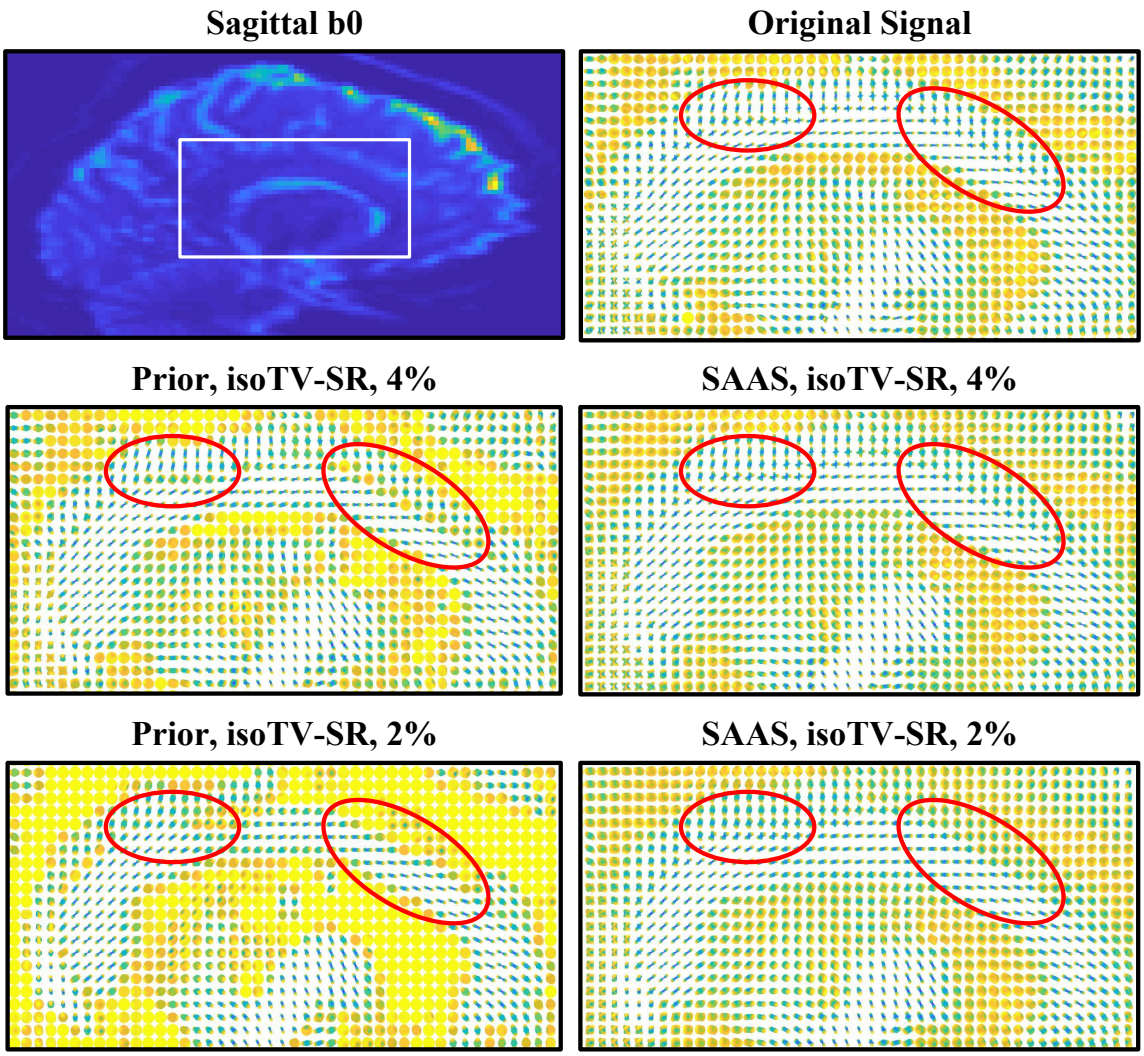


Figure 4.7: Reconstruction of corpus callosum in the sagittal view comparing (SAAS) and (Prior) (k, q)-CS. Top left: whole brain $b0$ image with ROI. Top right: ODFs in ROI estimated from fully sampled original signal. Middle: ODFs estimated from reconstructed signal with only 4% of the total (k, q) measurements, keeping 20% k-space samples and 20% q-space samples (51 grad dirs). Bottom: repeated with 2% of the total (k, q) measurements, keeping only 10% q-space samples (25 grad dirs). (Prior) is unable to reconstruct crossing fibers and sets many voxels to zero (yellow) while (SAAS) maintains accurate reconstruction at these very low sampling rates.

4.6.3 Generalization to Multiple Subjects

In this section, we demonstrate how well our proposed (k, q)-CS method performs on a large number of real HARDI subjects. The first question we wish to address is how well will our method generalizes to multiple subjects under a pre-chosen set of parameters tuned from a single subject. In the previous set of experiments, the two major parameters used for reconstruction were λ, which controls the trade-off between sparsity and reconstruction accuracy in the LASSO framework in (4.8), and the percentage of subsampling in k- and q-space.

The important parameter λ was tuned for each individual reconstruction experiment by cycling through a range of λ values and selecting the one that gave the lowest reconstruction error. A value of λ that is too large will produce a solution that is too sparse and not as accurate while one that is too small will produce a solution that is accurate with respect to the sampled data points but does not accurately model the unseen data points. Selecting the best λ from a range can be a time-consuming process, especially for large data sets, and is not a realistic method in a real CS experiment on newly acquired data. Similarly, the rate of subsampling is not something that can be optimized retrospectively. Therefore we wish to understand how well our method generalizes in a large study of HARDI subjects given a value of λ and a fixed subsampling rate, that have been tuned on a similar single subject.

We use 46 subjects from the Hippocampal Connectivity Project (HCP) at the Center for Imaging of Neurodegenerative Diseases at the University of California San

Francisco. Each subject was acquired on a Siemens 4T scanner (128 gradient directions, 3 b0 values, FOV: 192, number of slices: 26, resolution: 1.5 mm isotropic, b-value: 1400 s/mm^2, TR/TE: 3500/86). For our experiments we reduced the reconstruction to the interior of the white matter, a $50 \times 50 \times 20$ volume of interest for each of the 46 subjects.

We first took one of the 46 subjects from the HCP study and tuned the parameter λ and varied the percentage of subsampling to identify the optimal parameters for this subject. Figure 4.8 shows the reconstruction errors for each regime. We can see that for a specific range of λ between $1.4^{-3.4}$ and $1.4^{-6.4}$, the reconstruction errors reach a minimum for most sampling rates. The lowest amount of sampling was 6% (20% q-space, 30% k-space). Each sampling rate after that was equal between k and q, where, for example, 9% total subsampling is comprised of 30% in k and 30% in q subsampling and 16% total subsampling is 40% in k and 40% in q. For this experiment we choose to use 6% subsampling with an optimal parameter of $\lambda = 1.4^{-6.4}$, which is the minimum of the blue curve in Figure 4.8 and see how these parameter choices fare with regards to the reconstruction of the 45 other HCP subjects. The base 1.4 and exponents were tuned relative to this subject. These results are displayed in Figure 4.9. For comparison of reconstruction error, we also chose 16% subsampling, which may be a more usual and conservative choice for subsampling rates.

The second question we address in these experiments is a comparison between separable and non-separable sensing schemes as described in Section 4.6.1. Given

fixed separable and non-separable sampling schemes we compare how they perform on the same 45 real HARDI subjects also in Figure 4.9

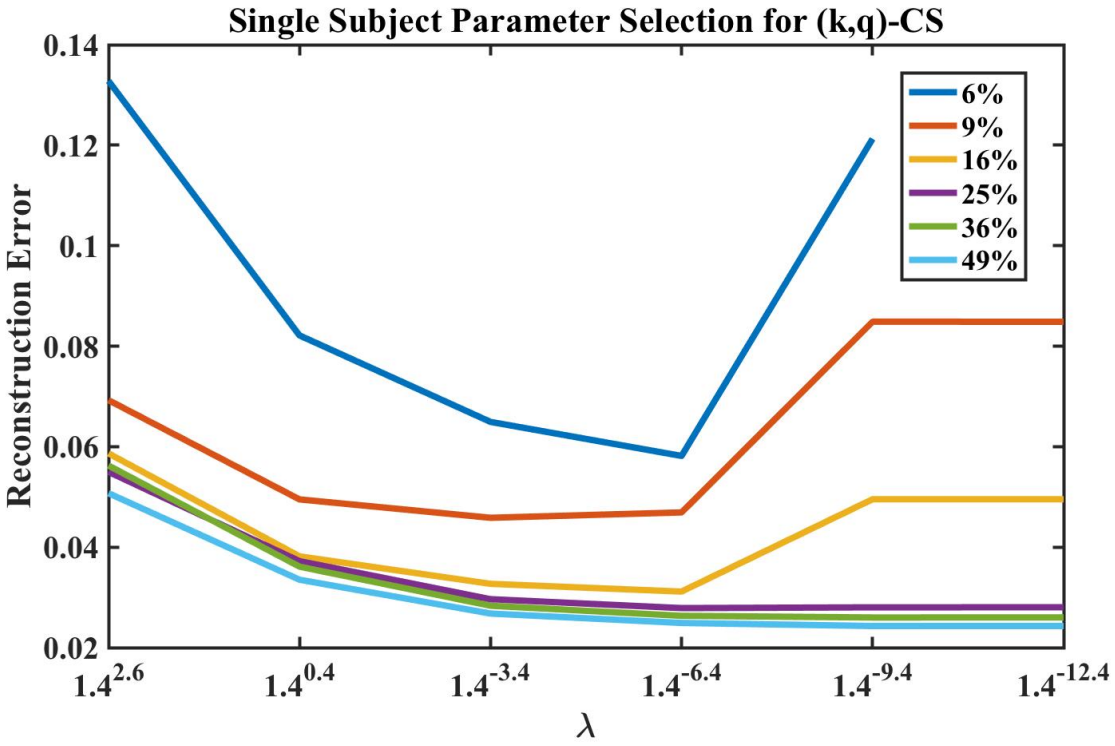


Figure 4.8: Reconstruction results for a single subject of the HCP HARDI data for various values of parameter λ and sampling rates. We choose the minimum sampling rate that gives us good reconstruction errors and the λ value that gives us the minimum. For this reason we choose 6% subsampling with $\lambda = 1.4^{-6.4}$ (minimum of blue curve) to be used for the remaining 45 subjects in the HCP data.

We can see that for the majority of the subjects, the error is consistent with that of the subject on which the parameters were tuned with an average of 0.051 error. Of the 45 subjects, 4 in a reconstruction error of about 0.2 which may be a result of the parameters being non-optimal for those subjects. In addition, for these subjects we compared the effects of separable vs. non-separable sensing and it can be seen that non-separable sensing results in a consistently lower error for the majority of

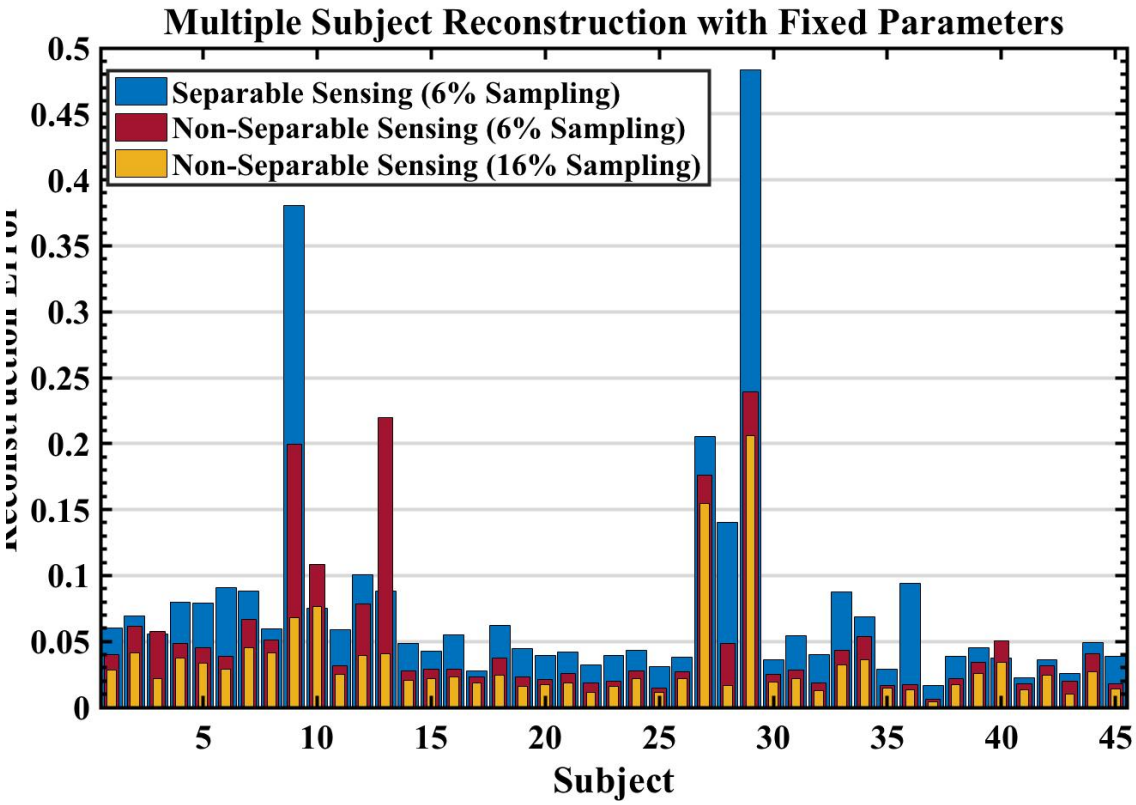


Figure 4.9: Reconstruction error of subjects from HCP HARDI study using the parameters tuned from a single subject: $\lambda = 1.4^{-6.4}$ at 6% total subsampling selected from Fig. 4.8. Also in comparison are two different sensing schemes, separable vs. non-separable sensing. We can see consistently accurate reconstruction errors for the vast majority of subjects. In addition it is evident that in most subjects non-separable sensing outperforms separable sensing. The four outliers may be a result of suboptimal λ for those subjects.

the subjects by an average of 0.0325. Only 3 of the subjects (Subjects 10, 13, and 40) have a lower error for separable sensing, and of those, one (Subject 13) may be due to suboptimal parameter tuning. These results indicate that non-separable sensing is superior to separable sensing and that our proposed SAAS (k, q)-CS can be used effectively and consistently to reconstruct real HARDI data at low subsampling schemes. In comparison to the 16% subsampling, we can see an improvement for every subject, especially Subjects 9 and 13.

4.7 Conclusion

In this work, we have proposed a unified (k, q)-CS model for dMRI that naturally exploits sparsity in the joint spatial-angular domain. The main goal of this chapter was to demonstrate the performance gains of CS using our joint model compared to state-of-the-art frameworks which combine k-CS and q-CS in an additive way. We have shown that we can achieve more accurate signal reconstructions with a greater reduction of measurements than state-of-the-art (k, q)-CS models, on the order of 2-6% of the original data. In addition, we have derived a novel Kronecker extension of FISTA to efficiently solve this large-scale optimization by exploiting the separability of Kronecker dictionaries. Though we experimented on single-shell HARDI, our proposed framework is general to any dMRI acquisition protocol with an appropriate choice of sensing and angular dictionary .

To make a concrete comparison of (k, q)-CS methods, we chose fixed sparsifying transforms/dictionaries and (k, q) sensing schemes and used a spatial-angular *analysis-synthesis* model to match that of state-of-the-art formulations. We have shown that our method generalizes well with a pre-tuned set of parameters to a number of test subjects. Furthermore, we have demonstrated the superior performance of non-separable sensing over separable sensing for a large number of real HARDI subjects.

We hope that the underlying framework for (k, q)-CS in this thesis may lead to increased levels of dMRI acceleration for greater practical usability in the future. Additional work will be needed to investigate the relationship between sampling in

(k, q)-space as a function of acquisition time. In the next chapter, we will expand upon our choices of sparsifying dictionaries by learning joint spatial-angular dictionaries directly from dMRI data with the aim of further increasing the sparsity levels we can achieve over that of fixed dictionaries and push the boundaries of (k, q)-CS subsampling.

Chapter 5

Spatial-Angular Dictionary Learning

5.1 Introduction

Until now, we have developed sparse coding and compressed sensing methods for dMRI which have utilized fixed analytic dictionaries such as spherical harmonics or spherical ridgelets for the angular domain and wavelets, curvelets or total variation for the spatial domain. While these dictionaries have demonstrated high levels of sparsification, they have been developed for many general signals and are not specific to the structure of dMRI data. In response, the idea of *learning* dictionaries directly from data has been shown to naturally produce sparser reconstruction which is beneficial for a number of applications in sparse coding, compressed sensing and other

signal processing tasks like denoising. This field, known as *dictionary learning* (see Chapter 2.2.3 for an overview), has gained increasing attention in signal processing with many advances in representing complex signal structures including many forms of medical images.

For dMRI, dictionary learning has been applied to learn angular q-space dictionaries from sets of voxel-wise training examples of q-space signals. However, the state of the art does not consider learning spatial dictionaries. The first major contribution in this chapter is to extend angular dictionary learning to also learn spatial dictionaries following the joint spatial-angular representation developed in our previous chapters. This framework fits within a subfield of dictionary learning known as *separable*, *Kronecker* or *tensor* dictionary learning due to the separability of the dictionaries.

In addition, we address a major limitation of current dictionary learning methods which is an absence of global optimality guarantees due to the non-convexity of the dictionary learning problem. The second major contribution in this chapter is a new framework for learning separable dictionaries which comes equipped with theoretical and algorithmic guarantees for global optimality. This stems from formulations in matrix factorization from the work of [142–145]. Before developing our novel separable dictionary learning method with global optimality, we review the state of the art first in angular dictionary learning for dMRI and then in separable dictionary learning.

5.2 State of the Art in Dictionary Learning

5.2.1 Angular Dictionary Learning for dMRI

Angular dictionary learning has been proposed for many dMRI applications including de-noising and compressed sensing and aims to learn an angular dictionary $\Gamma \in \mathbb{R}^{G \times r}$ by solving the following problem:

$$\min_{\Gamma, W} \; g(W) \;\; \text{s.t.} \;\; \frac{1}{2}||\Gamma W - Y||_F^2 \leq \epsilon, \;\; ||\Gamma_i||_2 \leq 1 \;\; \forall \, i = 1 \ldots r, \tag{5.1}$$

where $Y = [s_1, \ldots, s_T] \in \mathbb{R}^{G \times T}$ is comprised of T training examples of angular signals, $s_t \in \mathbb{R}^G$ taken from various white matter voxels in a brain image, for example, $W = [a_1, \ldots, a_T]$ are the corresponding coefficients for each training example and $g(A)$ is a function that induces properties in W, usually the L_0 or L_1 norm or the nuclear norm for a low-rank solution. Furthermore, the size of the dictionary r may be variable based on the application.

This (5.1) is an instance of the classical dictionary learning problem discussed in Section 2.2.3 and as such, there have been a multitude of works that aim to solve (5.1). Some propose alternative models like parametric dictionary learning [88,90,106, 109, 110, 123], which learn parameters from fixed diffusion models, Bayesian learning

[146, 147], manifold learning [111], and dictionary learning directly on undersampled data for CS [96, 105, 108, 146, 148]. Each of these various methods learns a purely angular q-space dictionary used to reconstruct signals in each voxel. While some of the mentioned work impose spatial coherence between angular dictionaries learned in neighboring voxels [123], they have not attempted to learn spatial dictionaries for dMRI.

In this Chapter, we learn spatial and angular dictionaries for dMRI jointly in order to provide sparser representations of dMRI than the analytical dictionaries used previously in this thesis. To the best of our knowledge, only the work of [149] has proposed to learn both spatial and angular dictionaries. However, their method restricts to the case of local, non-separable dictionaries on spatial-angular patches. We discuss the state of the art in separable dictionary learning next.

In our previous chapters we have demonstrated that sparse coding with separable dictionaries over the spatial and angular domain provides sparser reconstructions than the traditional angular sparse coding with important applications in (k, q)-CS. Thus, we should be able to provide sparser reconstructions with spatial-angular dictionary learning than angular dictionary learning. To the best of our knowledge, only one other recent work [149] learns spatial and angular dictionaries but restricts to the case of local, non-separable spatial-angular patch dictionaries, for denoising applications.

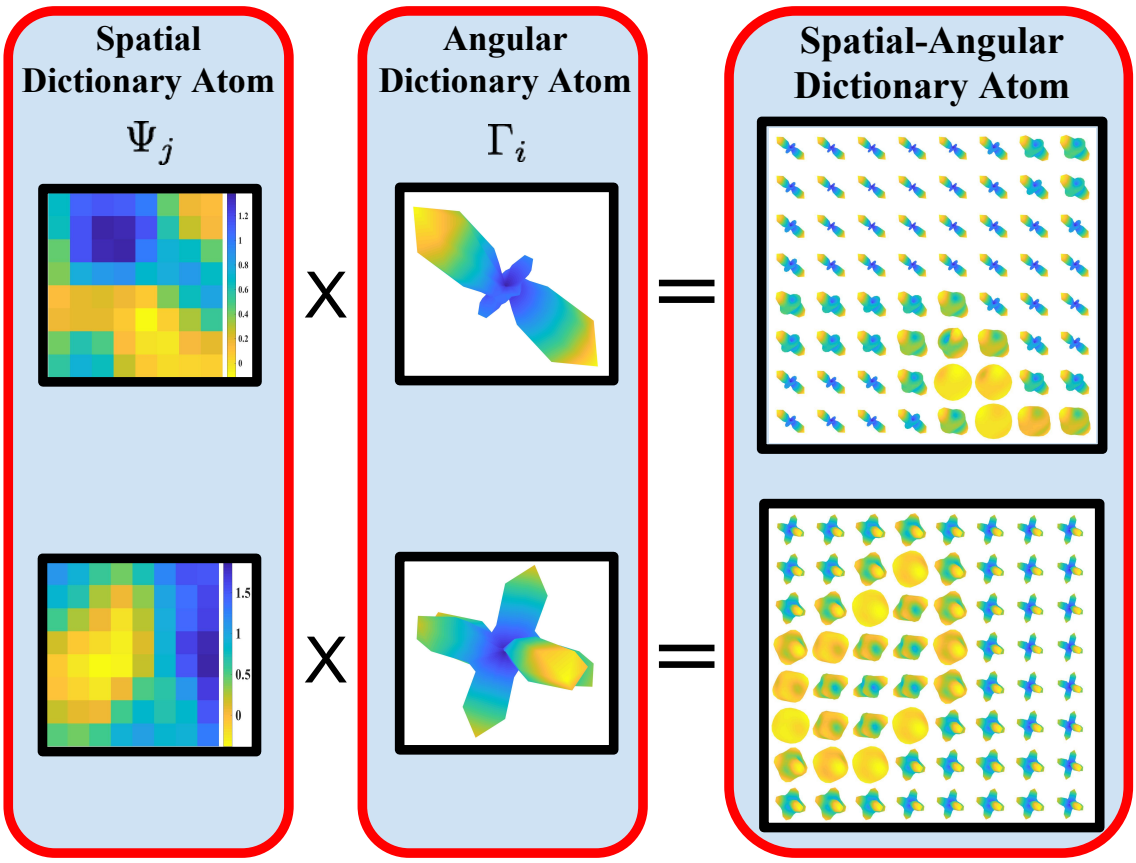


Figure 5.1: Spatial-Angular dictionary examples for 8×8 patches learned from phantom data. With a single spatial-angular atom, we can model complex fiber configurations in a given spatial neighborhood suggesting we can very sparsely represent dMRI data with by learning joint spatial-angular dictionaries.

5.2.2 Separable Dictionary Learning

To motivate the separable dictionary learning problem we recall that with our proposed spatial-angular representation of dMRI signals, we have $S = \Gamma C \Psi^\top$ where $S \in \mathbb{R}^{G \times V}$ with G gradient directions and V voxels, $\Gamma \in \mathbb{R}^{G \times r_1}$ is an angular dictionary and $\Psi \in \mathbb{R}^{V \times r_2}$ is a spatial dictionary, and $C \in \mathbb{R}^{r_1 \times r_2}$ are the joint spatial-angular coefficients. Here we use variables r_1 and r_2 for the number of atoms of each dictionary (instead of N_Γ and N_Ψ previously) because the sizes of the dictionaries may not be

known *a priori* and could change within the optimization. Now, recall that for the case of angular dictionary learning, we have T angular training examples $s_t \in \mathbb{R}^G$, each with coefficients $a_t \in \mathbb{R}^r$, such that $s_t = \Gamma a_t$ for all t. Now, for spatial-angular dictionary learning, we will have T spatial-angular training examples $S_t \in \mathbb{R}^{G \times V}$, each with coefficients $C_t \in \mathbb{R}^{r_1 \times r_2}$, such that $S_t = \Gamma C_t \Psi^\top$ for all t. In this setting, separable dictionary learning aims to solve the following problem:

$$\min_{\Gamma, \Psi, \{C_t\}} \frac{1}{2} \sum_{t=1}^{T} ||\Gamma C_t \Psi^\top - S_t||_F^2 + \lambda ||C_t||_1 \;\; \text{s.t.} \;\; ||\Gamma_i||_2 \leq 1, ||\Psi_j||_2 \leq 1 \;\; \forall (i,j). \quad (5.2)$$

Learning separable dictionaries via (5.2) (and alternative formulations such as its multidimensional tensor generalization or low-rank regularization) has been studied previously in the literature. The work of [130, 150, 151] solve variations of (5.2) using conjugate gradient methods over smooth manifolds. In terms of tensors, [135, 152], resort to solving alternatively each mode of the tensor as the usual vector dictionary learning problem after n-mode unfolding, which loses the computational gain of maintaining a tensor structure. The work of [153–155] use decompositions such as Tucker, Kruskal-Factor and tensor SVDs, while [156] considers a dictionary as the sum of Kronecker products. Finally, [157, 158] propose to solve low-rank variations (5.2).

As we recall for (2.54), one key difficulty in dictionary learning is the lack of guarantees of global optimality due to the non-convexity of the joint optimization over the

dictionary and coefficients. This issue is especially difficult for separable dictionary learning because the number of variables to jointly optimize over increases from two to three or more. To the best of our knowledge, none of the aforementioned work on separable dictionary learning come equipped with guarantees for global optimality, and so their solutions may correspond to a local minimum or saddle point and may also heavily depend on initialization. The main contribution of this work is a new framework for separable dictionary learning with guarantees of global optimality. To do this, we expand upon theoretical work on matrix factorization [142, 145] which has been applied previously to provide theoretical guarantees to the original dictionary learning problem (2.54).

In the next Section 5.3 we will show how the classic dictionary learning problem (5.1) or (2.54) can be posed as a matrix factorization problem and in this setting, how guarantees of global optimality can be achieved based on the results of [145]. Then in Section 5.4 we will extend the theories of matrix factorization to handle the case of separable dictionary learning (5.2) and provide guarantees of global optimality for this problem. With these theoretical guarantees, we will then present a novel algorithm to find global minimum of the separable dictionary learning problem. Finally, in Section 5.5 we will test our methods on the application of denoising diffusion magnetic resonance imaging data.

5.3 Background

5.3.1 Dictionary Learning as Matrix Factorization

The general problem of matrix factorization is concerned with finding factors D and W, such that a data matrix Y can be approximated by a matrix $X = DW$. Naturally, the dictionary learning problem (2.54) can be thought of in this way. In [142–145] the authors develop a general matrix factorization framework for a number of applications including the dictionary learning problem. The key insight is an equivalence relation between the non-convex factorized problem with respect to the factors D and W and a convex problem with respect to X, which allows to obtain guarantees of global optimality for (D, W).

First, the non-convex matrix factorization problem can be written as:

$$\min_{D,W} \ell(Y, DW) + \lambda\Theta(D, W), \tag{5.3}$$

where ℓ is a data fidelity term or loss that measures the error between the original signal Y and the reconstruction $X = DW$, and Θ is a regularizer on the factors D and W which promotes particular properties relevant to the problem. For the dictionary learning problem (2.54), $\ell(Y, DW) = \frac{1}{2}||DW - Y||_F^2$. Furthermore it can be shown that the constraints $||D_i||_2 \leq 1$ can be combined with the sparsity term $||W||_1$ to get $\Theta(D, W) = \sum_{i=1}^{r} ||D_i||_2 ||W_i^\top||_1$, where $W_i^\top \in \mathbb{R}^T$ is the i^{th} row of W.

Then (5.3) is an equivalent problem formulation of (2.54). The goal of recasting the dictionary learning problem as a matrix factorization problem is to relate (5.3), which is non-convex with respect to D and W, to a convex problem with respect to X.

5.3.2 Global Optimality for Matrix Factorization

To derive conditions for the global optimality, [145] first impose the regularizer to be of the specific form $\Theta(D, W) = \sum_i^r \theta(D_i, W_i^\top)$ where θ is a rank-1 regularizer that must satisfy the following properties:

Definition 7. [from [145]] A function $\theta : \mathbb{R}^N \times \mathbb{R}^T \to \mathbb{R}_+ \cup \infty$ is said to be a **rank-1 regularizer** if

1. $\theta(u, v)$ is positively homogeneous with degree 2, i.e. $\theta(\alpha u, \alpha v) = \alpha^2 \theta(u, v)\ \forall \alpha \geq 0,\ \forall (u, v)$.

2. $\theta(u, v)$ is positive semi-definite, i.e. $\theta(0, 0) = 0$ and $\theta(u, v) \geq 0\ \forall (u, v)$.

3. For any sequence (u_n, v_n) such that $||u_n v_n^\top|| \to \infty$, we have that $\theta(u_n, v_n) \to \infty$.

It is easy to show that the choice of $\theta(u, v) = ||u||_2 ||v||_1$ fits this definition. Another example of θ satisfying Definition 7 that can be used for dictionary learning is $\theta(u, v) = \|u\|_2(\|v\|_2 + \alpha\|v\|_1)$ which promotes column regularization in u and v and also sparsity in v as analyzed in [142].

Now, in order to connect the non-convex (5.3) with a convex problem with respect to matrix X, we introduce a related regularizer $\Omega_\theta(X)$ which depends on θ:

Definition 8. [from [145]] Given a rank-1 regularizer θ that satisfies the conditions of Definition 7, the **matrix factorization regularizer** $\Omega_\theta : \mathbb{R}^{N\times T} \to \mathbb{R}_+ \cup \infty$ is defined as:

$$\Omega_\theta(X) \equiv \inf_{r\in\mathbb{N}_+} \inf_{D,W} \sum_{i=1}^{r} \theta(D_i, W_i^\top) \;\text{ s.t. }\; DW = X. \tag{5.4}$$

If the infimum is achieved for some D, W and r then we say that DW is an optimal factorization of X.

It is important to note that the number of dictionary atoms r becomes an important variable of finding an optimal matrix factorization in this definition. As a motivating example for the origin of Ω_θ, when $\theta(D_i, W_i^\top) = ||D_i||_2||W_i^\top||_2$, $\Omega_\theta(X)$ becomes the variational definition of the nuclear norm:

$$||X||_* \equiv \inf_{r\in\mathbb{N}_+} \inf_{D,W} \sum_{i=1}^{r} ||D_i||_2||W_i^\top||_2 \;\text{ s.t. }\; DW = X. \tag{5.5}$$

From the results of [145], $\Omega_\theta(X)$ is a gauge function (and even a norm if θ is symmetric, i.e. $\theta(-u, v) = \theta(u, v)$ or $\theta(u, -v) = \theta(u, v)$ for all u, v), which leads to the new convex optimization problem with respect to X:

$$\min_X \ell(Y, X) + \lambda\Omega_\theta(X). \tag{5.6}$$

Since (5.6) is convex, a local minimum is guaranteed to be global, $\hat{X}$. The question answered in [145] is then how to relate a local minimum $(\tilde{D}, \tilde{W})$ of the non-convex (5.3)

to a global minimum of the convex (5.6), $\hat{X}$, and when, if ever, can we say something about global minimum $(\hat{D}, \hat{W})$ of (5.3). First, it is evident that (5.6) provides a global lower bound of (5.3) because Ω_θ is the infimum of Θ and $\ell(Y, X) = \ell(Y, DW)$. The main result is then that under certain conditions local minima $(\tilde{D}, \tilde{W})$ of the non-convex (5.3) are optimal factorizations of X, such that $\hat{X} = \tilde{D}\tilde{W}$. In other words, given a local solution $(\tilde{D}, \tilde{W})$ to (5.3), we can write a matrix $X = \tilde{D}\tilde{W}$ and under certain conditions, it turns out that the matrix X is a global minimum of (5.6), i.e. $X \equiv \hat{X}$. Therefore, $(\tilde{D}, \tilde{W})$ is in fact a global minimum of (5.3), $(\hat{D}, \hat{W})$. We restate this main theorem of [145] here:

Theorem 9. *[from [145]] Given a function $\ell(S, X)$ that is convex and once differentiable w.r.t. X, a rank-1 regularizer θ that satisfies the conditions in Definition 7, with constants $r \in \mathbb{N}_+$, and $\lambda > 0$, local minima $(\tilde{D}, \tilde{W})$ of (5.3) are globally optimal if $(\tilde{D}_i, \tilde{W}_i^\top) = (0, 0)$ for some $i \in [r]$. Moreover, $\hat{X} = \tilde{D}\tilde{W}$ is a global minima of (5.6) and $\tilde{D}\tilde{W}$ is an optimal factorization of $\hat{X}$.*

Since θ is general, this matrix factorization can be applied to many problems such as low-rank, non-negative matrix factorization, sparse PCA as well as the desired dictionary learning. However, one important downside for the application of dictionary learning is that the choices of θ stated above are not well suited to checking the criteria of Theorem 9 in practice. In particular verifying if a point is stationary or a local minimum is remains difficult. Therefore, finding globally optimal solutions for classical dictionary learning still remains a challenging problem. In the next section

we will extend the results of [145] for the more complex structured *separable* dictionary learning. Moreover, given a certain choice of regularizer, we do not run into the same issues as for the classical dictionary learning problem when finding globally optimal solutions in practice.

5.4 Proposed Separable Dictionary Learning with Global Optimality

In this section, we extend the formulation of dictionary learning as a matrix factorization problem to that of separable dictionary learning (5.2) as a *tensor* factorization problem to give guarantees of global optimality.

5.4.1 Separable Dictionary Learning as Tensor Factorization

Similar to matrix factorization, tensor factorization is concerned with finding factors that decompose a n-tensor $\underline{S} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_n}$. There are two main types of tensor decompositions: rank-1 decomposition, where each factor $f_i \in \mathbb{R}^{N_i}$ is a vector such that $\underline{X} = f_1 \otimes f_2 \otimes \cdots \otimes f_n$, where $\otimes$ is the tensor outer product, and the Tucker decomposition, in which there is a core n-tensor $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \cdots \times r_n}$ and matrix factors $F_i \in \mathbb{R}^{N_i \times r_i}$ such that $\underline{X} = \underline{C} \times_1 F_1 \times_2 F_2 \cdots \times_n F_n$, where $\times_n$ stands for matrix

multiplication on the n^{th} dimension of the core tensor $\underline{C}$. (See [159] for a review of tensor decomposition.)

To link tensor factorization to the separable dictionary problem (5.2), consider again that we wish to find dictionaries Γ and Ψ such that signal $S_t = \Gamma C_t \Psi^\top$ for each training example $t = 1, \ldots, T$. By stacking each matrix $S_t \in \mathbb{R}^{G \times V}$ and $C_t \in \mathbb{R}^{r_1 \times r_2}$ as slices of 3-tensors $\underline{S} \in \mathbb{R}^{G \times V \times T}$ and $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}$ using tensor multiplication notation, this is equivalent to writing $\underline{S} = \underline{C} \times_1 \Gamma \times_2 \Psi$. The goal of tensor factorization is then to find factors Γ and Ψ such that the data tensor $\underline{S}$ can be approximated by $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$.

For notation, all tensors will be written with an underline, e.g. the 3-tensor $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}$. To index the tensor $\underline{C}$, all 2D slices (matrices) will be written with an upper case letter and a single index, e.g. $C_t \in \mathbb{R}^{r_1 \times r_2}$ or $C_i \in \mathbb{R}^{r_2 \times T}$ or $C_j \in \mathbb{R}^{r_1 \times T}$. Next, 1D vectors of $\underline{C}$ will be written with an upper case letter and two indices, e.g. $C_{i,j} \in \mathbb{R}^T$ or $C_{i,t} \in \mathbb{R}^{r_2}$ or $C_{j,t} \in \mathbb{R}^{r_1}$. Finally, single elements (scalars) of $\underline{C}$ will be written in lowercase with three indices, $c_{i,j,t}$.

Similar to non-convex matrix factorization problem (5.3) in the previous section, the non-convex tensor factorization problem can be stated as:

$$\min_{\Gamma, \Psi, \underline{C}} \{ f(\Gamma, \Psi, \underline{C}) \equiv \ell(\underline{S} \ , \ \underline{C} \times_1 \Gamma \times_2 \Psi) + \lambda \Theta(\Gamma, \Psi, \underline{C}) \}. \tag{5.7}$$

To link to the separable dictionary learning problem (5.2), we set $\ell(\underline{S}, \underline{C} \times_1 \Gamma \times_2$

$\Psi) = \frac{1}{2}||\underline{C} \times_1 \Gamma \times_2 \Psi - \underline{S}||_F^2 = \frac{1}{2}\sum_{t=1}^{T}||\Gamma C_t \Psi^\top - S_t||_F^2$. Then by combining the constraints $||\Gamma_i||_2 \leq 1$ and $||\Psi_i||_2 \leq 1$ with the sparse regularizer $\sum_{t=1}^{T}||C_t||_1 = ||\underline{C}||_1 = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}||C_{i,j}||_1$ we can write

$$\Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}||\Gamma_i||_2||\Psi||_2||C_{i,j}||_1. \tag{5.8}$$

It can be easily shown that with this choice of Θ, (5.7) is an equivalent reformulation of the separable dictionary learning problem (5.2). Now, as before, we wish to link stationary points of the non-convex $f(\Gamma, \Psi, \underline{C})$ with a global minimum of a convex function with respect to $\underline{X}$.

5.4.2 Global Optimality for Tensor Factorization

To develop the theories of global optimality for separable dictionary learning we begin by extending Definitions 7 and 8 from Section **??**. First, we will consider a regularizer in (5.7) of the form $\Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\theta(\Gamma_i, \Psi_j, C_{i,j})$ where θ satisfies the following conditions:

Definition 10. A function $\theta : \mathbb{R}^G \times \mathbb{R}^V \times \mathbb{R}^T \to \mathbb{R}_+ \cup \infty$ is said to be a **rank-1 regularizer** if

1. θ is positively homogeneous of degree 3, i.e. $\theta(\alpha\gamma, \alpha\psi, \alpha c) = \alpha^3\theta(\gamma, \psi, c)$ $\forall \alpha \geq 0,\ \forall(\gamma, \psi, c)$.

2. θ is positive semi-definite and $\theta(\gamma, \psi, c) > 0$ for any (γ, ψ, c) s.t. $\gamma \otimes \psi \otimes c \neq 0$.

3. For any sequence (γ_n, ψ_n, c_n) such that $||\gamma_n \otimes \psi_n \otimes c_n|| \to \infty$, we have $\theta(\gamma_n, \psi_n, c_n) \to \infty$.

Proposition 2. *The regularizer $\theta(\gamma, \psi, c) = ||\gamma||_2||\psi||_2||c||_1$ satisfies Definition 10.*

Proof. For $\alpha \geq 0$, $\theta(\alpha\gamma, \alpha\psi, \alpha c) = ||\alpha\gamma||_2||\alpha\psi||_2||\alpha c||_1 = \alpha^3||\gamma||_2||\psi||_2||c||_1 = \alpha^3\theta(\gamma, \psi, c)$, positively homogeneous of degree 3. Because θ is a multiplication of norms, θ is positive semi-definite and positive for any triple (γ, ψ, c) with $\gamma \neq 0, \psi \neq 0, c \neq 0$. Finally, the last property is trivially verified since $\|\cdot\|$ and θ are equivalent norms on $\mathbb{R}^{G\times V\times T}$. □

Then, similarly to Definition 8, we define the related regularizer for tensor $\underline{X}$:

Definition 11. Given a rank-1 regularizer θ that satisfies the conditions of Definition 10, the **tensor factorization regularizer** $\Omega_\theta : \mathbb{R}^{G\times V\times T} \to \mathbb{R}_+ \cup \infty$ is defined as:

$$\Omega_\theta(\underline{X}) := \inf_{r_1,r_2\in\mathbb{N}_+} \inf_{\substack{\Gamma\in\mathbb{R}^{G\times r_1}\\ \Psi\in\mathbb{R}^{V\times r_2}\\ \underline{C}\in\mathbb{R}^{r_1\times r_2\times T}}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \underline{X}. \quad (5.9)$$

If the infimum is achieved for some $(\Gamma, \Psi, \underline{C})$ and $r_1, r_2 \in \mathbb{N}_+$ then we say that $\underline{C} \times_1 \Gamma \times_2 \Psi$ is an optimal factorization of $\underline{X}$.

Proposition 3. *Given regularizer θ that satisfies the properties of Definition 10, the tensor factorization regularizer $\Omega_\theta(\underline{X})$ satisfies the following properties:*

1. *$\Omega_\theta(0) = 0$ and $\Omega_\theta(\underline{X}) > 0$ $\forall \underline{X} \neq 0$.*

2. $\Omega_\theta(\alpha\underline{X}) = \alpha\Omega_\theta(\underline{X})\ \forall\alpha \geq 0\ \forall\underline{X}$.

3. $\Omega_\theta(\underline{X} + \underline{Y}) \leq \Omega_\theta(\underline{X}) + \Omega_\theta(\underline{Y})\ \forall(\underline{X}, \underline{Y})$.

4. If θ *is symmetric in* γ,ψ *or* c, *then* $\Omega_\theta(-\underline{X}) = \Omega_\theta(\underline{X})\ \forall\underline{X}$ *and* Ω_θ *is a norm.*

5. The infimum of $\Omega_\theta(\underline{X})$ *in* (11) *can be achieved with finite* r_1 *and* r_2.

Proof. We will assume, in a first phase, that the infimum in (5.9) can be achieved for finite r_1 and r_2 (which is proved in the last point below) and drop the minimization in r_1 and r_2 to lighten the derivations.

1. First, since $\theta(\gamma, \psi, c) \geq 0\ \forall(\gamma, \psi, c)$, we have that $\Omega_\theta(\underline{X}) \geq 0\ \forall\underline{X}$. Then, the infimum $\Omega_\theta(0) = 0$ can be achieved by taking $(\Gamma, \Psi, \underline{C}) = (0, 0, 0)$. If $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$ and $\underline{X} \neq 0$, we can write equivalently $\underline{X} = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \Gamma_i \otimes \Psi_j \otimes C_{i,j}$ and there exists i_0, j_0 such that $\Gamma_{i_0} \neq 0, \Psi_{j_0} \neq 0, C_{i_0,j_0} \neq 0$ and thus $\Omega_\theta(\underline{X}) \geq \theta(\Gamma_{i_0}, \Psi_{j_0}, C_{i_0,j_0}) > 0$ thanks to the second property in Proposition 2.

2. With the substitution $(\bar{\Gamma}, \bar{\Psi}, \bar{\underline{C}}) := (\alpha^{-1/3}\Gamma, \alpha^{-1/3}\Psi, \alpha^{-1/3}\underline{C})$ and using the pos-

itive homogeneity of θ,

$$\begin{aligned}
\Omega_\theta(\alpha\underline{X}) &= \inf_{\Gamma,\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \alpha\underline{X} \\
&= \inf_{\Gamma,\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } (\alpha^{-1/3})\underline{C} \times_1 (\alpha^{-1/3})\Gamma \times_2 (\alpha^{-1/3})\Psi = \underline{X} \\
&= \inf_{\bar{\Gamma},\bar{\Psi},\bar{\underline{C}}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\alpha^{1/3}\bar{\Gamma}_i, \alpha^{1/3}\bar{\Psi}_j, \alpha^{1/3}\bar{C}_{i,j}) \text{ s.t. } \bar{\underline{C}} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\
&= \inf_{\bar{\Gamma},\bar{\Psi},\bar{\underline{C}}} \alpha \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\bar{\Gamma}_i, \bar{\Psi}_j, \bar{C}_{i,j}) \text{ s.t. } \bar{\underline{C}} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\
&= \alpha\Omega_\theta(\underline{X}).
\end{aligned}$$

3. Let $\underline{X} = \underline{C}_x \times_1 \Gamma_x \times_2 \Psi_x$ and $\underline{Y} = \underline{C}_y \times_1 \Gamma_y \times_2 \Psi_y$ be two ϵ-optimal factorizations such that $\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\theta(\Gamma_{x_i}, \Psi_{x_j}, C_{x_{i,j}}) = \Omega_\theta(\underline{X}) + \epsilon$ and same for $\underline{Y}$. Now we construct $\Gamma = [\Gamma_x,\ \Gamma_y]$, $\Psi = [\Psi_x,\ \Psi_y]$, and $\underline{C} = \begin{bmatrix} \underline{C}_x & 0 \\ 0 & \underline{C}_y \end{bmatrix}$ such that $\underline{X} + \underline{Y} = \underline{C} \times_1 \Gamma \times_2 \Psi$. Then $\Omega_\theta(\underline{X} + \underline{Y}) \leq \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} ||\Gamma_i||_2||\Psi_j||_2||C_{i,j}||_1 \leq \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} ||\Gamma_{x_i}||_2||\Psi_{x_j}||_2||C_{x_{i,j}}||_1 + ||\Gamma_{y_i}||_2||\Psi_{y_j}||_2||C_{y_{i,j}}||_1 = \Omega_\theta(\underline{X}) + \Omega_\theta(\underline{Y}) + 2\epsilon$, due to the triangle inequality of the respective norms. Taking $\epsilon \to 0$ completes the triangle inequality.

4. Assuming $\theta(-\gamma, \psi, c) = \theta(\gamma, \psi, c)$, (as is true for $-\psi$ or $-c$) and setting $\bar{\Gamma} := -\Gamma$,

$$\begin{aligned}
\Omega_\theta(-\underline{X}) &= \inf_{\Gamma,\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = -\underline{X} \\
&= \inf_{\Gamma,\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 -\Gamma \times_2 \Psi = \underline{X} \\
&= \inf_{\bar{\Gamma},\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(-\bar{\Gamma}_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \bar{\Gamma} \times_2 \Psi = \underline{X} \\
&= \inf_{\bar{\Gamma},\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\bar{\Gamma}_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \bar{\Gamma} \times_2 \Psi = \underline{X} \\
&= \Omega_\theta(\underline{X}).
\end{aligned}$$

5. In order to show that there exists a global minimum with finite r_1 and r_2 in the definition of Ω_θ, we begin by remarking that we have the following alternative definition:

$$\Omega_\theta(\underline{X}) = \inf_{r_1,r_2 \in \mathbb{N}_+} \inf_{\substack{\Gamma \in \mathbb{R}^{G \times r_1} \\ \Psi \in \mathbb{R}^{V \times r_2} \\ \underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}}} \|\underline{C}\|_1 \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \underline{X} \text{ and } \|\Gamma_i\|_2 \|\Psi_j\|_2 \leq 1 \, \forall (i,j) \tag{5.10}$$

where $\|\underline{C}\|_1$ is the sum of the absolute values of all the entries in the tensor $\underline{C}$. This follows simply from the homogeneity of the regularizer θ. Now, we

introduce a second polar function defined as follows:

$$\widetilde{\Omega}_\theta(\underline{X}) := \inf_{r\in\mathbb{N}_+} \inf_{\substack{\Gamma\in\mathbb{R}^{G\times r}\\ \Psi\in\mathbb{R}^{V\times r}\\ \Lambda\in\mathbb{R}^{T\times r}}} \|\Lambda\|_1 \;\text{ s.t. }\; \sum_{i=1}^{r} \Gamma_i\otimes\Psi_i\otimes\Lambda_i = \underline{X} \text{ and } \|\Gamma_i\|_2\|\Psi_i\|_2 \leq 1 \;\forall i \tag{5.11}$$

where again $\|\Lambda\|_1$ is the sum of the absolute values of the entries of the matrix Λ. The important point of note is that the minimization in $\widetilde{\Omega}_\theta$ is exactly the same as the one appearing in (5.10) with the extra constraints that $r_1 = r_2 = r$ and that $\underline{C}$ is a slice by slice diagonal tensor, i.e for all $t = 1,\dots,T$ $\underline{C}_{\cdot,\cdot,t}$ is a diagonal matrix (with Λ_t then corresponding to the diagonal of this matrix). Therefore, one immediately obtains that $\widetilde{\Omega}_\theta(\underline{X}) \geq \Omega_\theta(\underline{X})$. This is in fact an equality. Indeed, let $r_1, r_2 \in \mathbb{N}_+$ and $\Gamma, \Psi, \underline{C}$ such that $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$ and $\|\Gamma_i\|_2\|\Psi_j\|_2 \leq 1$ for all (i,j). Then, define $r = r_1 r_2$ and consider the lexicographic ordering of pairs $l : \{1,\dots,r_1\}\times\{1,\dots,r_2\} \to \{1,\dots,r\}$. Setting $\tilde{\Gamma} \in \mathbb{R}^{G\times r}$ such that $\tilde{\Gamma}_{l(i,j)} = \Gamma_i$, $\tilde{\Psi} \in \mathbb{R}^{V\times r}$ such that $\tilde{\Psi}_{l(i,j)} = \Psi_j$ and $\Lambda_{l(i,j),t} = \underline{C}_{i,j,t}$, we see that $\underline{X} = \sum_{l=1}^{r} \tilde{\Gamma}_l \otimes \tilde{\Psi}_l \otimes \Lambda_l$, $\|\tilde{\Gamma}_l\|_2\|\tilde{\Psi}_l\|_2 \leq 1$ for all $l = 1,\dots,r$ and $\|\Lambda\|_1 = \|\underline{C}\|_1$. Consequently, any value of the minimization problem (5.10) can be obtained by (5.11) thanks to the previous transformation and we get that $\widetilde{\Omega}_\theta(\underline{X}) = \Omega_\theta(\underline{X})$.

We now only need to show that a global minimum in (5.11) can be achieved with a finite r, which will give a global minimum of (5.10) with $r_1 = r_2 = r$. We follow an argument similar to the one presented in [145] that we briefly recap.

First, for the same reason as above, we have

$$\widetilde{\Omega}_\theta(\underline{X}) := \inf_{r\in\mathbb{N}_+} \inf_{\substack{\Gamma\in\mathbb{R}^{G\times r}\\ \Psi\in\mathbb{R}^{V\times r}\\ \Lambda\in\mathbb{R}^{T\times r}}} \sum_{i=1}^{r} \|\Gamma_i\|_2\|\Psi_i\|_2\|\Lambda_i\|_1 \;\text{ s.t. }\; \sum_{i=1}^{r} \Gamma_i\otimes\Psi_i\otimes\Lambda_i = \underline{X}. \quad (5.12)$$

Let $\Theta \subset \mathbb{R}^{G\times V\times T}$ defined by $\Theta = \{\underline{X} :\ \exists(\gamma,\psi,\lambda)/\ \underline{X} = \gamma\otimes\psi\otimes\lambda \text{ and } \|\gamma\|_2\|\psi\|_2\|\lambda\|_1 \leq 1\}$. With the same reasoning as [145], we know that $\widetilde{\Omega}_\theta$ is equivalent to the following gauge function on the convex hull of Θ:

$$\widetilde{\Omega}_\theta(\underline{X}) = \inf\{\mu :\ \mu \geq 0, \underline{X} \in \mu\text{conv}(\Theta)\}$$

Now since Θ and thus $\text{conv}(\Theta)$ are compact sets, the previous infimum over μ is achieved for a certain $\mu^* \geq 0$. Then $\underline{X} \in \mu^*\text{conv}(\Theta)$ and from Caratheodory's theorem, we know that any point in $\text{conv}(\Theta)$ can be written as a finite convex combination of a most $G\times V\times T$ elements in Θ. In other words, there exists $(\Gamma_i^*,\Psi_i^*,\Lambda_i)_{i=1,\ldots,r}$ with $r \leq G\times V\times T$ such that $\underline{X} = \mu^*\sum_{i=1}^r \Gamma_i^*\otimes\Psi_i^*\otimes\Lambda_i = \sum_{i=1}^r \Gamma_i^*\otimes\Psi_i^*\otimes\Lambda_i^*$ with $\Lambda_i^* = \mu^*\Lambda_i$ for all i. Now since $\mu^* = \widetilde{\Omega}_\theta(\underline{X})$, we obtain eventually $\sum_{i=1}^r \|\Gamma_i^*\|_2\|\Psi_i^*\|_2\|\Lambda_i^*\|_1 = \widetilde{\Omega}_\theta(\underline{X})$ which is thus a global minimum with finite r.

□

By definition, satisfying the first three properties show that Ω_θ is gauge function, and properties 2 and 3 show that Ω_θ is convex. Furthermore, for our choice of

$\theta(\gamma, \psi, c) = ||\gamma||_2||\psi||_2||c||_1)$, $\Omega_\theta(\underline{X})$ is a norm. Then, with respect to $\underline{X}$, we have the convex problem:

$$\min_{\underline{X}}\{F(\underline{X}) \equiv \ell(\underline{S}, \underline{X}) + \lambda\Omega_\theta(\underline{X})\}, \tag{5.13}$$

where F is a global lower bound for f. The next theorem, an extension of Theorem 9, which relates the non-convex f (5.7) to the convex F (5.13), is the main result of this manuscript.

Theorem 12. *Given a function $\ell(\underline{S}, \underline{X})$ that is convex and once differentiable w.r.t. $\underline{X}$, a rank-1 regularizer θ that satisfies the conditions in Definition 10, and constants $r_1, r_2 \in \mathbb{N}_+$, and $\lambda > 0$, any local minima $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{\underline{C}})$ of $f(\Gamma, \Psi, \underline{C})$ in (5.7) is globally optimal if there exists (i, j) such that $(\tilde{\Gamma}_i, \tilde{\Psi}_j) = (0, 0)$ and for all t, $(\tilde{C}_{i,t}, \tilde{C}_{j,t}) = (0, 0)$. Moreover, $\hat{\underline{X}} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$ is a global minimum of $F(\underline{X})$ in (5.13) and $\tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$ is an optimal factorization of $\hat{\underline{X}}$ in (5.9).*

In order to prove Theorem 12, we first note that $\hat{\underline{X}}$ is a global minimum of $F(\underline{X})$ if and only if $-\frac{1}{\lambda}\nabla_{\underline{X}}\ell(\underline{S}, \hat{\underline{X}}) \in \partial\Omega_\theta(\hat{\underline{X}})$ since we have a convex function. Therefore, we must first characterize the subgradient $\partial\Omega_\theta(\underline{X})$ which is the subject of the following lemma.

Lemma 13. *The subgradient $\partial\Omega_\theta(\underline{X})$ is given by:*

$$\left\{\underline{W} : \langle \underline{W}, \underline{X}\rangle = \Omega_\theta(\underline{X}) \quad \textit{and} \quad \sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c)\ \forall(\gamma, \psi, c)\right\}. \tag{5.14}$$

Proof. Since Ω_θ is convex, by Fenchel duality, $\underline{W} \in \partial\Omega_\theta$ if and only if $\langle \underline{W}, \underline{X}\rangle =$

$\Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W})$ where Ω_θ^* is the Fenchel dual of Ω_θ given by $\Omega_\theta^*(\underline{W}) \equiv \sup_{\underline{Z}} \langle \underline{W}, \underline{Z} \rangle - \Omega_\theta(\underline{Z})$. From the definition of $\Omega_\theta(\underline{Z})$ we can expand the dual as

$$\begin{aligned}
\Omega_\theta^*(\underline{W}) &= \sup_{r_1,r_2} \sup_{\Gamma,\Psi,\underline{C}} \langle \underline{W}, \underline{Z} \rangle - \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \underline{Z} \\
&= \sup_{r_1,r_2} \sup_{\Gamma,\Psi,\underline{C}} \langle \underline{W}, \underline{C} \times_1 \Gamma \times_2 \Psi \rangle - \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \\
&= \sup_{r_1,r_2} \sup_{\Gamma,\Psi,\underline{C}} \sum_{t=1}^{T} \langle \Gamma^T W_t \Psi, C_t \rangle - \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \\
&= \sup_{r_1,r_2} \sup_{\Gamma,\Psi,\underline{C}} \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \left(\sum_{t=1}^{T} c_{i,j,t} \Gamma_i^\top W_t \Psi_j \right) - \theta(\Gamma_i, \Psi_j, C_{i,j})
\end{aligned} \tag{5.15}$$

If there exists (γ, ψ, c) such that $\sum_{t=1}^{T} c_t \gamma^\top W_t \psi > \theta(\gamma, \psi, c)$, we can see that $\Omega_\theta^*(\underline{W}) = \infty$ by considering $(\alpha\gamma, \alpha\psi, \alpha c)$ as $\alpha \to \infty$ and using the positive homogeneity of θ.

Now let $\underline{W} \in \partial\Omega_\theta$. Then $\Omega_\theta^*(\underline{W}) < +\infty$ and thus, from the previous argument, we have that $\sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c)\ \forall(\gamma, \psi, c)$. This also implies that all the terms in the summation in (5.15) will be non-positive leaving the supremum to be 0, achieved with $(\Gamma, \Psi, \underline{C}) = (0, 0, \underline{0})$. It follows that $\Omega_\theta^*(\underline{W}) = 0$ and consequently $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$.

Conversely, if $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$ and $\sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c)\ \forall(\gamma, \psi, c)$ then, reasoning as previously, we see that $\Omega_\theta^*(\underline{W}) = 0$ which implies $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W})$ and thus $\underline{W} \in \partial\Omega_\theta(\underline{X})$. □

Next, using the characterization of $\partial\Omega_\theta(\underline{X})$ in Lemma 13, we identify when a factorization $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$ is optimal, i.e. when a point $(\Gamma, \Psi, \underline{C})$ achieves the

infimum of $\Omega_\theta(\underline{X})$ in (5.9), in the following corollary.

Corollary 14. *For factorization $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$, if there exists $\underline{W}$ such that $\langle \underline{W}, \underline{X} \rangle = \Theta(\Gamma, \Psi, \underline{C})$ and $\sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c)\ \forall(\gamma, \psi, c)$, then $\underline{W} \in \partial\Omega_\theta(\underline{X})$ and $\underline{C} \times_1 \Gamma \times_2 \Psi$ is an optimal factorization of $\underline{X}$, i.e. it achieves the infimum of $\Omega_\theta(\underline{X})$.*

Proof. By contradiction, assume $\underline{W} \notin \partial\Omega_\theta(\underline{X})$. Then $\langle \underline{W}, \underline{X} \rangle < \Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W}) = \Omega_\theta(\underline{X})$ because $\sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c)\ \forall(\gamma, \psi, c)$, implies $\Omega_\theta^*(\underline{W}) = 0$ as in the proof of Lemma 13. Then, from our assumption, $\langle \underline{W}, \underline{X} \rangle = \Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) < \Omega_\theta(\underline{X})$ which violates the definition of $\Omega_\theta(\underline{X})$ being the infimum, producing a contradiction. Therefore, $\underline{W} \in \partial\Omega_\theta(\underline{X})$. Now, since $\underline{W} \in \partial\Omega_\theta(\underline{X})$, by Lemma 13, $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$, which implies $\Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) = \Omega_\theta(\underline{X})$ thus showing that $\underline{C} \times_1 \Gamma \times_2 \Psi$ achieves the infimum of $\Omega_\theta(\underline{X})$ and is an optimal factorization of $\underline{X}$. □

Finally, with Lemma 13 and Corollary 14 we can now prove Theorem 12:

Proof of Theorem 12. From (5.13), we know $\hat{\underline{X}} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$ is a global minimum of $F(\underline{X})$ if and only if $-\frac{1}{\lambda} \nabla_{\underline{X}} \ell(\underline{S}, \hat{\underline{X}}) \in \partial\Omega_\theta(\hat{\underline{X}})$. Notice $-\frac{1}{\lambda} \nabla_{\underline{X}} \ell(\underline{S}, \hat{\underline{X}})$ can be written in terms of its slices as $\sum_{t=1}^{T} -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \hat{X}_t) = \sum_{t=1}^{T} -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)$. To prove that $\hat{\underline{X}} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$ is a global minimum and an optimal factorization of $\hat{\underline{X}}$, from Corollary 14, it suffices to show two conditions:

1. $\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{t=1}^{T} \tilde{c}_{i,j,t} \tilde{\Gamma}_i^\top (-\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)) \tilde{\Psi}_j = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j})$

2. $\sum_{t=1}^{T} c_t \gamma^\top (-\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)) \psi \leq \theta(\gamma, \psi, c) \; \forall (\gamma, \psi, c)$

To show condition 1, let $\Gamma_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3} \tilde{\Gamma}$ and $\Psi_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3} \tilde{\Psi}$ and $\underline{C}_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3} \underline{\tilde{C}}$. Since $(\tilde{\Gamma}, \tilde{\Psi}, \underline{\tilde{C}})$ is a local minimum, there exists $\delta > 0$ such that for all $\epsilon \in (0, \delta)$ we have

$$\sum_{t=1}^{T} \ell(S_t, \Gamma_{1\pm\epsilon} C_{t_{1\pm\epsilon}} \Psi_{1\pm\epsilon}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta((1 \pm \epsilon)^{1/3} \tilde{\Gamma}_i, (1 \pm \epsilon)^{1/3} \tilde{\Psi}_j, (1 \pm \epsilon)^{1/3} \tilde{C}_{i,j}) \tag{5.16}$$

$$= \sum_{t=1}^{T} \ell(S_t, (1 \pm \epsilon) \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) + \lambda (1 \pm \epsilon) \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) \tag{5.17}$$

$$\geq \sum_{t=1}^{T} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}). \tag{5.18}$$

Rearranging the last inequality gives

$$\frac{-1}{\lambda \epsilon} [\sum_{t=1}^{T} \ell(S_t, (1 \pm \epsilon) \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) - \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)] \leq \pm \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}). \tag{5.19}$$

Taking the limit as $\epsilon \searrow 0$ gives the directional derivative:

$$\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) \leq \sum_{t=1}^{T} \langle \frac{-1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top), \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top \rangle \leq \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) \tag{5.20}$$

which implies equality. Rearranging the inner product gives Condition 1.

Next, to show condition 2, we use the assumption that there exists (i, j) such that $(\tilde{\Gamma}_i, \tilde{\Psi}_j) = (0, 0)$ and for all t, $(\tilde{C}_{i,t}, \tilde{C}_{j,t}) = (0, 0)$. Without loss of generality let the

last column pair of $(\tilde{\Gamma}, \tilde{\Psi})$ be zero and the last columns and rows of $\underline{\tilde{C}}$ be zero for all t. Then, given (γ, ψ, c), let $\Gamma_\epsilon = [\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_{r_1-1}, \epsilon^{1/3}\gamma]$ and $\Psi_\epsilon = [\tilde{\Psi}_1, \ldots, \tilde{\Psi}_{r_2-1}, \epsilon^{1/3}\psi]$ and

$$C_{t_\epsilon} = \begin{bmatrix} \underline{\tilde{c}}_{1,1,t} & \cdots & \underline{\tilde{c}}_{1,r_2-1,t} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \underline{\tilde{c}}_{r_1-1,1,t} & \cdots & \underline{\tilde{c}}_{r_1-1,r_2-1,t} & 0 \\ 0 & \ldots & 0 & \epsilon^{1/3} c_t \end{bmatrix} \quad \forall t. \tag{5.21}$$

Now, for all $\epsilon > 0$ sufficiently small we have

$$\begin{aligned} &\sum_{t=1}^{T} \ell(S_t, \Gamma_\epsilon C_{t_\epsilon} \Psi_\epsilon^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) + \lambda\theta(\epsilon^{1/3}\gamma, \epsilon^{1/3}\psi, \epsilon^{1/3}c) = \\ &\sum_{t=1}^{T} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top + \epsilon c_t \gamma\psi^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) + \epsilon\lambda\theta(\gamma, \psi, c) \geq \\ &\sum_{t=1}^{T} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}), \end{aligned} \tag{5.22}$$

where the first equality follows from θ being positively homogeneous and the last inequality from the fact that $\underline{\tilde{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$ is assumed to be a local minimum of $F(\Gamma, \Psi, \underline{C})$ and by choosing ϵ small enough. Therefore, by rearranging the inequality we arrive at:

$$\frac{-1}{\lambda\epsilon}[\sum_{t=1}^{T} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top + \epsilon c_t\gamma\psi^\top) - \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top)] \leq \theta(\gamma, \psi, c) \tag{5.23}$$

Since $\ell(S_t, X_t)$ is differentiable with respect to X_t, taking the limit as $\epsilon \searrow 0$, the

directional derivative in the direction of $c_t\gamma\psi^\top$ gives us

$$\sum_{t=1}^{T}\langle\frac{-1}{\lambda}\nabla_{X_t}\ell(S_t,\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top),c_t\gamma\psi^\top\rangle\leq\theta(\gamma,\psi,c) \tag{5.24}$$

$$\implies\sum_{t=1}^{T}c_t\gamma^\top(\frac{-1}{\lambda}\nabla_{X_t}\ell(S_t,\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top)\psi\leq\theta(\gamma,\psi,c), \tag{5.25}$$

which proves Condition 2. This together with Condition 1 proves Theorem 12. □

The results of Theorem 12 are true for any local minima of f. But, in general, local descent methods (e.g. gradient descent) are only guaranteed to converge to a stationary point at most and therefore arriving at local minima of f is not guaranteed, i.e. it may be possible to reach a saddle point. However, with our particular choice of θ, we can derive sufficient conditions for global optimality of any point $(\Gamma,\Psi,\underline{C})$.

Corollary 15. *Let $\theta(\gamma,\psi,c)=||\gamma||_2||\psi||_2||c||_1$. A point $(\tilde{\Gamma},\tilde{\Psi},\tilde{\underline{C}})$ is a global minimum of $f(\Gamma,\Psi,\underline{C})$ if and only if it satisfies the following conditions:*

1. $\sum_{t=1}^{T}\tilde{c}_{i,j,t}\tilde{\Gamma}_i^\top(-\frac{1}{\lambda}\nabla_{X_t}\ell(S_t,\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top))\tilde{\Psi}_j=||\tilde{\Gamma}_i||_2||\tilde{\Psi}_j||_2||\tilde{C}_{i,j}||_1\ \forall\ (i,j)$

2. $\max_{1\leq t\leq T}\sigma_{max}(-\frac{1}{\lambda}\nabla_{X_t}\ell(S_t,\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top))\leq 1$

where σ_{max} is the maximum singular value.

Proof. First, we know that to be a global minimum, a point must first satisfy first-order optimality for f. Noting that $\theta(\tilde{\Gamma}_i,\tilde{\Psi}_i,\tilde{C}_{i,j})=\|\tilde{\Gamma}_i\|_2\|\tilde{\Psi}_j\|_2\sum_{t=1}^{T}|\tilde{c}_{i,j,t}|$ and writ-

ing the first-order optimality conditions on the coefficients $\tilde{c}_{i,j,t}$, we obtain that:

$$0 = \tilde{\Gamma}_i^\top \nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top)\tilde{\Psi}_j + \lambda \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \text{ sign}(\tilde{c}_{i,j,t}) \tag{5.26}$$

for all $i = 1, \ldots, r_1$, $j = 1, \ldots, r_2$ and $t = 1, \ldots, T$ if $\tilde{c}_{i,j,t} \neq 0$. Multiplying by $\tilde{c}_{i,j,t}$ then leads, in all cases, to:

$$0 = \tilde{c}_{i,j,t}\tilde{\Gamma}_i^\top \nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top)\tilde{\Psi}_j + \lambda \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 |\tilde{c}_{i,j,t}|. \tag{5.27}$$

Now, summing over t gives for all i, j:

$$\sum_{t=1}^{T} \tilde{c}_{i,j,t}\tilde{\Gamma}_i^\top (-\frac{1}{\lambda}\nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top))\tilde{\Psi}_j = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \sum_{t=1}^{T} |\tilde{c}_{i,j,t}| = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \|\tilde{C}_{i,j}\|_1. \tag{5.28}$$

Therefore, if a point satisfies condition 1 than it is a stationary point.

Next, from Theorem 12, we know that for a stationary point to be a global minimum we need to check that the following condition is satisfied:

$$\sum_{t=1}^{T} c_t \gamma^\top (-\frac{1}{\lambda}\nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top))\psi \leq \theta(\gamma, \psi, c) \ \forall(\gamma, \psi, c). \tag{5.29}$$

For simplicity let $W_t := -\frac{1}{\lambda}\nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top)$. With our choice of θ, this condition becomes:

$$\sum_{t=1}^{T} c_t \gamma^\top W_t \psi \leq ||\gamma||_2 ||\psi||_2 ||c||_1 \ \ \forall(\gamma, \psi, c). \tag{5.30}$$

Now, normalizing each variable by its respective norm, such that $\hat{\gamma} = \gamma/||\gamma||_2, \hat{\psi} = \psi/||\psi||_2, \hat{c}_t = c_t/||c||_1$, the previous condition becomes

$$\sum_{t=1}^{T} \hat{c}_t \hat{\gamma}^\top W_t \hat{\psi} \leq 1 \quad \forall \hat{\gamma} \neq 0, \hat{\psi} \neq 0, \hat{c} \neq 0, \tag{5.31}$$

which we can equivalently state as:

$$\sup_{||\hat{\gamma}||_2 = ||\hat{\psi}||_2 = ||\hat{c}||_1 = 1} \sum_{t=1}^{T} \hat{c}_t \hat{\gamma}^\top W_t \hat{\psi} \leq 1. \tag{5.32}$$

Now, with respect to $\hat{c}$, since $||\hat{c}||_1 = 1$, the supremum of a linear combination can be attained by choosing $\hat{c}_{t_*} = 1$ and $c_t = 0$ for $t \neq t_*$ with t_* such that $\sup_{||\hat{\gamma}||_2 = ||\hat{\psi}||_2 = 1} \hat{\gamma}^\top W_{t_*} \hat{\psi} = \max_t \{\sup_{||\hat{\gamma}||_2 = ||\hat{\psi}||_2 = 1} \hat{\gamma}^\top W_t \hat{\psi}\}$. Therefore, (5.32) is equivalent to:

$$\max_{1 \leq t \leq T} \{ \sup_{||\hat{\gamma}||_2 = ||\hat{\psi}||_2 = 1} \hat{\gamma}^\top W_t \hat{\psi} \} \leq 1, \tag{5.33}$$

and, with σ_{max} denoting the largest singular value of the corresponding matrix, this is the same as:

$$\max_{1 \leq t \leq T} \sigma_{max}(W_t) \leq 1, \tag{5.34}$$

which shows condition 2. Thus, if a point satisfies conditions 1 and 2 than it is a global minimum of f. Conversely, if a point is a global minimum of f, then it satisfies first-order optimality which is equivalent to condition 1. Next, a global minimum of f will be a minimum of F. Therefore, the point will be contained in $\partial\Omega_\theta(\underline{X})$ and

condition 2 of Corollary 14 will be satisfied which is equivalent to condition 2 in this proof. Thus, we have shown the equivalency of this proof. □

Using the results of Corollary 15, we can devise an algorithm to find a global minimum of of the separable dictionary learning problem by first finding a stationary point of (5.2) and then checking if it satisfies condition 2 in Corollary 15. A logical next question of this routine is what happens if the stationary point does not satisfy (2). In [145], the authors demonstrate that by appending a column of zeros to the dictionary $\tilde{D}$ and a row of zeros to the coefficients $\tilde{W}$, they are guaranteed to continue in a descent direction. Therefore, the algorithm will consist of iterating between local descent and global optimality check and appending the resulting stationary point. In this way, the optimal size of the dictionary, r, is learned through the process.

For separable dictionary with two dictionaries, we have two size parameters r_1 and r_2. Therefore, we have an additional option to augment one or both of the dictionaries to proceed in descent directions. Based on the application or *a prior* knowledge of the relative dictionary sizes, we have the opportunity to schedule the increments of r_1 and r_2. The following proposition proves that In the next section we will formalize a novel algorithm that alternates between these two sub-routines until convergence.

5.4.3 Algorithm to Reach Global Minimum

Now that we are equipped with conditions to guarantee global minima for separable dictionary learning, we will outline an algorithm to reach a globally optimal solution. This involves alternating between two main sub-routines: 1) local descent to reach a stationary point with fixed number of atoms r_1 and r_2 in the dictionaries, and 2) a check for global optimality via Corollary 15. Note that since we consider the particular choice of regularizer $\theta(\gamma, \psi, c) = \|\gamma\|_2\|\psi\|_2\|c\|_1$, the global optimality check only amounts to verifying that a stationary point satisfies condition 2 in Corollary 15. If by the end of the local descent we have not reached a globally optimal solution, then we can find a global descent direction by adding additional atoms to the dictionaries. Algorithm 11 describes this general meta-algorithm in more detail and refers to each sub-routine discussed in the following sections.

Algorithm 11 Meta-Algorithm: Local Descent and Global Optimality Check

```
Initialize dictionaries with set number of atoms.
while not globally optimal do
  while objective residual > ϵ do
    descent to local minimum via Algorithm 12
  end while
  if Condition 2 is satisfied then
    solution is globally optimal
  else
    update dictionaries via Algorithm 14
  end if
end while
```

5.4.3.1 Proximal Gradient Descent to Stationary Point

In this section, we provide an algorithm to find a stationary point of the separable dictionary learning problem with fixed sizes for the dictionaries. We state the problem again here:

$$\min_{\Gamma,\Psi,\underline{C}} \frac{1}{2}\sum_{t=1}^{T} ||\Gamma C_t \Psi^\top - S_t||_F^2 + \lambda \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} ||\Gamma_i||_2||\Psi_j||_2||C_{i,j}||_1. \tag{5.35}$$

For an optimization problem of the form $\min_x\{\ell(x)+\lambda\Theta(x)\}$, where ℓ is differentiable and Θ is non-differentiable, proximal gradient descent is a common algorithm to arrive at a stationary points, i.e. local minima or saddle points. The general updates for proximal gradient descent follow:

$$x^{k+1} = \text{prox}_{\tau\lambda\Theta(\cdot)}(x^k - \tau\nabla\ell), \tag{5.36}$$

where $\text{prox}_{\tau\lambda\Theta(\cdot)}(y) = \arg\min_x\{\frac{1}{2\tau\lambda}||x-y||_2^2+\Theta(x)\}$. To solve (5.35), we apply a proximal gradient descent step to each variable while holding the rest constant. The local descent algorithm is outlined in Algorithm 12. Recall $\ell(\Gamma,\Psi,\underline{C}) = \frac{1}{2}\sum_{t=1}^{T}||\Gamma C_t\Psi^\top -$

$S_t||_F^2$. We derive the update for each variable as:

$$\Gamma_i^{k+1} = \text{prox}_{\xi_i^k||\cdot||_2}(\Gamma_i^k - \xi_i^k[\nabla_{\Gamma^k}\ell]_i) \tag{5.37}$$

$$c_{i,j,t}^{k+1} = \text{prox}_{\kappa_{i,j}^k|\cdot|}(c_{i,j,t}^k - \kappa_{i,j}^k[\nabla_{C_t^k}\ell]_{i,j}) \tag{5.38}$$

$$\Psi_j^{k+1} = \text{prox}_{\pi_j^k||\cdot||_2}(\Psi_j^k - \pi_j^k[\nabla_{\Psi^k}\ell]_j). \tag{5.39}$$

where the proximal operators for $||\cdot||_2$ and $|\cdot|$ can be written in closed form:

$$\text{prox}_{\tau||\cdot||_2}(x) = \begin{cases} (1 - \frac{\tau}{||x||_2})x & \text{for } ||x||_2 \geq \tau \\ 0 & \text{otherwise} \end{cases}, \tag{5.40}$$

$$\text{prox}_{\tau|\cdot|}(\alpha) = \max(0, \alpha - \tau) - \max(0, -\alpha - \tau), \tag{5.41}$$

for $x \in \mathbb{R}^N$, $\alpha \in \mathbb{R}$ and $\tau \geq 0$. Then

$$\nabla_\Gamma \ell = \sum_{t=1}^{T} (\Gamma C_t \Psi^\top - S_t)\Psi C_t^\top \tag{5.42}$$

$$\nabla_{C_t} \ell = \Gamma^\top (\Gamma C_t \Psi^\top - S_t)\Psi \tag{5.43}$$

$$\nabla_\Psi \ell = \sum_{t=1}^{T} (\Psi C_t^\top \Gamma^\top - S_t^\top)\Gamma C_t. \tag{5.44}$$

Finally, $\xi_i, \kappa_{i,j}$, and π_j are constants composed of the other fixed variables in θ, i.e.

$$\xi_i := \lambda \sum_{t=1}^{T} \sum_{j=1}^{r_2} |c_{i,j,t}| ||\Psi_j||_2 / L_\Gamma \tag{5.45}$$

$$\kappa_{i,j} := \lambda ||\Gamma_i||_2 ||\Psi_j||_2 / L_{C_t} \tag{5.46}$$

$$\pi_j := \lambda \sum_{t=1}^{T} \sum_{i=1}^{r_1} |c_{i,j,t}| ||\Gamma_i||_2 / L_\Psi, \tag{5.47}$$

where the parameters $1/L_\Gamma, 1/L_{C_t}$, and $1/L_\Psi$ correspond to the step sizes in the proximal gradient descent.

In general, to determine an appropriate step-size τ, it has been shown that convergence is guaranteed if $\tau \leq \frac{1}{L}$, where L is the Lipschitz constant of ∇l:

$$||\nabla \ell(x^{(1)}) - \nabla \ell(x^{(2)})||_2 \leq L ||x^{(1)} - x^{(2)}||_2. \tag{5.48}$$

In our this setting, we can calculate (or at least bound) the Lipschitz constants with

respect to, $L_\Gamma, L_{\underline{C}_t}$, and L_Ψ. For L_Γ we have indeed:

$$\begin{aligned}
||\nabla_\Gamma \ell(\Gamma^{(1)}) - \nabla_\Gamma \ell(\Gamma^{(2)})||_F &= ||\sum_{t=1}^{T} (\Gamma^{(1)} C_t \Psi^\top - S_t)\Psi C_t^\top - (\Gamma^{(2)} C_t \Psi^\top - S_t)\Psi C_t^\top||_F \\
&= ||\sum_{t=1}^{T} \Gamma^{(1)} C_t \Psi^\top \Psi C_t^\top - \Gamma^{(2)} C_t \Psi^\top \Psi C_t^\top||_F \\
&= ||\sum_{t=1}^{T} C_t \Psi^\top \Psi C_t^\top (\Gamma^{(1)} - \Gamma^{(2)})||_F \\
&\leq ||\sum_{t=1}^{T} C_t \Psi^\top \Psi C_t^\top||_F ||(\Gamma^{(1)} - \Gamma^{(2)})||_F \\
&= L_\Gamma ||(\Gamma^{(1)} - \Gamma^{(2)})||_F
\end{aligned}$$

where $L_\Gamma = ||\sum_{t=1}^{T} C_t \Psi^\top \Psi C_t^\top)||_F$ is thus an upper bound for the Lipschitz constant of $\nabla_\Gamma \ell$. Similarly for $\nabla_\Psi \ell$, we can take as the Lipschitz constant $L_\Psi = ||\sum_{t=1}^{T} C_t \Gamma^\top \Gamma C_t^\top||_F$. Then for $\nabla_{\underline{C}_t} \ell$, $L_{\underline{C}_t} = ||\Gamma^\top \Gamma||_F ||\Psi^\top \Psi||_F$. Lastly, the convergence of the descent can be accelerated through the standard Nesterov scheme as an extension of the Proximal Gradient Descent in Algorithm 13.

Algorithm 12 Proximal Gradient Descent

Initialize: $k = 0, \Gamma^0, \Psi^0, \underline{C}^0, \lambda, r_1, r_2$.
while error $> \epsilon$ **do**
  Update Γ^k via (5.37)
  Update $\underline{C}^k$ via (5.38)
  Update Ψ^k via (5.39)
  $k \to k + 1$
end while
return stationary point $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{\underline{C}})$

Algorithm 13 Proximal Gradient Descent with Nesterov Acceleration

Initialize: $k = 0, \check{\Gamma}^0, \check{\Psi}^0, \underline{\check{C}}^0, \lambda, r_1, r_2$.
while error $> \epsilon$ **do**
  $\Gamma_i^{k+1} = \text{prox}_{\xi_i^k||\cdot||_2}(\check{\Gamma}_i^k - \xi_i^k[\nabla_{\check{\Gamma}^k}\ell]_i)$
  $c_{i,j,t}^{k+1} = \text{prox}_{\kappa_{i,j}^k|\cdot|}(\check{c}_{i,j,t}^k - \kappa_{i,j}^k[\nabla_{\check{C}_t^k}\ell]_{i,j})$
  $\Psi_j^{k+1} = \text{prox}_{\pi_j^k||\cdot||_2}(\check{\Psi}_j^k - \pi_j^k[\nabla_{\check{\Psi}^k}\ell]_j)$.
  if $f(\Gamma^k, \Psi^k, \underline{C}^k) < f(\Gamma^{k-1}, \Psi^{k-1}, \underline{C}^{k-1})$ **then**
    $s_k = (1 + \sqrt{1 + 4s_{k-1}^2})/2$
    $\mu = (s_{k-1} - 1)/2$
    $\mu_\Gamma = \min(\mu, \sqrt{L_\Gamma^{k-1}/L_\Gamma^k})$
    $\mu_{C_t} = \min(\mu, \sqrt{L_{C_t}^{k-1}/L_{C_t}^k})\ \forall t$
    $\mu_\Psi = \min(\mu, \sqrt{L_\Psi^{k-1}/L_\Psi^k})$
    $\check{\Gamma}^{k+1} = \Gamma^k + \mu_\Gamma(\Gamma^k - \Gamma^{k-1})$
    $\check{C}_t^{k+1} = C_t^k + \mu_{C_t}(C_t^k - C_t^{k-1})$
    $\check{\Psi}^{k+1} = \Psi^k + \mu_\Psi(\Psi^k - \Psi^{k-1})$
  else
    $s_k = s_{k-1}$
    $\check{\Gamma}^{k+1} = \Gamma^{k-1}$
    $\check{C}_t^{k+1} = C_t^{k-1}\ \forall t$
    $\check{\Psi}^{k+1} = \Psi^{k-1}$
  end if
  $k \to k + 1$
end while
return stationary point $(\tilde{\Gamma}, \tilde{\Psi}, \underline{\tilde{C}})$

5.4.3.2 Global Optimality Check

Once proximal gradient descent reaches a local minimum via Algorithm 12, we check if the solution is a global minimum. By the results of the proof of Corollary 15, we can check if (2) is satisfied. If so, we have reached a global minimum and the algorithm stops. If not, by adding additional atoms to the dictionaries, we can try to escape from the local minimum we have reached and search for a global descent

direction. Following the discussions in [145] for matrix factorization involving the nuclear norm, by appending the locally optimal dictionaries with the singular vectors of the maximum SVD in (2), we are guaranteed to move in a global descent direction. First, let $t_* = \arg\max_t \sigma_{\max}(W_t)$. Then with $(\gamma_{t_*}, \psi_{t_*})$ the left and right singular vector pair corresponding to the maximum singular value of W_t over all t, we can update the locally optimal dictionaries $\tilde{\Gamma}$ and $\tilde{\Psi}$ by appending the last column $\Gamma = [\tilde{\Gamma}, \gamma_{t_*}]$ and $\Psi = [\tilde{\Psi}, \psi_{t_*}]$, with global step-size τ. Finally, $\underline{C}$ can be updated by appending the slice corresponding to the maximum singular value by

$$C_{t_*} = \begin{bmatrix} \tilde{C}_{t_*} & 0 \\ 0 & \tau \end{bmatrix} \tag{5.49}$$

and appending zero for all other slices. One important parameter to select is the global step-size τ. In our formulation, the optimal τ^* at each iteration can be found by solving:

$$\tau^* = \arg\min_{\tau} \frac{1}{2} \sum_{t}^{T} ||S_t - \hat{X}_t - \tau E_t||_F^2 + \lambda|\tau|, \tag{5.50}$$

where $E_{t_*} = \gamma_{t_*}\psi_{t_*}^\top$ and 0 for all other t. By vectorizing all tensors, (5.50) reduces to the simple proximal operator of the absolute value function with closed-form soft-thresholding solution.

Now, because of the separable form of this problem, we actually have the option to update just one of the two dictionaries, and not both simultaneously during each global check. In particular, if $\gamma_{t_*} \in \text{Span}(\tilde{\Gamma})$ then it is unnecessary to add this atom

to the dictionary. The same goes for ψ_{t_*}. Instead of checking these conditions *a posteriori*, we can check augmented criteria akin to (2) with the added constraint that $\gamma \in \text{Span}(\tilde{\Gamma})$ which by definition means that there exists an α such that $\gamma = \tilde{\Gamma}\alpha$. Using this we can make a change of variable in (2) as:

$$\sum_{t=1}^{T} c_t \alpha^\top \tilde{\Gamma}^\top W_t \psi \leq ||\tilde{\Gamma}\alpha||_2 ||\psi||_2 ||c||_1 \quad \forall(\alpha, \psi, c). \tag{5.51}$$

By noting that $||\tilde{\Gamma}\alpha||_2 \leq ||\tilde{\Gamma}||_2 ||\alpha||_2 = \sigma_{max}(\tilde{\Gamma})||\alpha||_2$, we can check the looser criteria:

$$\sum_{t=1}^{T} c_t \alpha^\top \tilde{\Gamma}^\top W_t \psi \leq \sigma_{max}(\tilde{\Gamma}) ||\alpha||_2 ||\psi||_2 ||c||_1 \quad \forall(\alpha, \psi, c). \tag{5.52}$$

Therefore, if (5.52) is violated then so is (5.51). We prefer to check (5.52) because of its simplicity to compute as follows. As before, normalizing $\hat{\psi} = \psi/||\psi||_2$, $\hat{c}_t = c_t/||c||_1$, and $\hat{\alpha}/||\alpha||_2$ give

$$\frac{1}{\sigma_{max}(\tilde{\Gamma})} \sum_{t=1}^{T} \hat{c}_t \hat{\alpha}^\top \tilde{\Gamma}^\top W_t \hat{\psi} \leq 1 \quad \forall(\hat{\alpha}, \hat{\psi}, \hat{c}) \tag{5.53}$$

$$\implies \sup_{\hat{\alpha}, \hat{\psi}, \hat{c}} \frac{1}{\sigma_{max}(\tilde{\Gamma})} \sum_{t=1}^{T} \hat{c}_t \hat{\alpha}^\top \tilde{\Gamma} W_t \hat{\psi} \leq 1 \quad \text{s.t.} \quad ||\hat{\alpha}||_2 = ||\hat{\psi}||_2 = ||\hat{c}||_1 = 1 \tag{5.54}$$

Algorithm 14 Global Optimality Check and Update

for $t = 1 \dots T$ **do**
  $\hat{X}_t = \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top$;
  $g_t = \sigma_{max}(-\hat{\tilde{\Gamma}}^\top(\hat{X}_t - S_t)/\lambda\sigma_{max}(\tilde{\Gamma}))$;
  $p_t = \sigma_{max}(-(\hat{X}_t - S_t)\tilde{\Psi}/\lambda\sigma_{max}(\tilde{\Psi}))$;
  $c_t = \sigma_{max}(-(\hat{X}_t - S_t)/\lambda)$;
end for
$g = \max_t g_t$;
$p = \max_t p_t$;
$c = \max_t c_t$;
if $g > 1$ and $g > p$ **then**
  Compute global step-size τ via (5.50)
  Update $\underline{C}$ and Ψ
else if $p > 1$ and $p > g$ **then**
  Compute global step-size τ via (5.50)
  Update Γ and $\underline{C}$
else if $c > 1$ **then**
  Compute global step-size τ via (5.50)
  Update Γ, $\underline{C}$ and Ψ
else
  $\Gamma^* = \tilde{\Gamma}; \underline{C}^* = \underline{\tilde{C}}; \Psi^* = \tilde{\Psi}$;
  $\hat{S} = \hat{X}$;
end if

Again we can check the stronger condition that:

$$\sup_{\hat{c}} \frac{1}{\sigma_{max}(\tilde{\Gamma})} \sum_{t=1}^{T} \hat{c}_t \sup_{\hat{\alpha},\hat{\psi}} \hat{\alpha}^\top \tilde{\Gamma}^\top W_t \hat{\psi} \leq 1 \quad \text{s.t.} \quad ||\hat{\alpha}||_2 = ||\hat{\psi}||_2 = ||\hat{c}||_1 = 1$$

$$\implies \sup_{||\hat{c}||_1=1} \frac{1}{\sigma_{max}(\tilde{\Gamma})} \sum_{t=1}^{T} \hat{c}_t \, \sigma_{max}(\tilde{\Gamma}^\top W_t) \leq 1$$

$$\implies \max_{1\leq t\leq T} \frac{1}{\sigma_{max}(\tilde{\Gamma})} \sigma_{max}(\tilde{\Gamma}^\top W_t) \leq 1 \tag{5.55}$$

If this inequality is violated, this implies that γ_{t_*}, the right singular vector of $\hat{\tilde{\Gamma}}^\top W_t$ corresponding to the maximum singular value could be appended to $\tilde{\Gamma}$ to give

a global descent direction. But because $\gamma_{t_*} \in \text{Span}(\tilde{\Gamma})$, it is not necessary to add it to find the descent direction. Therefore, we can just update Ψ and $\underline{C}$ as $\Psi = [\tilde{\Psi}, \psi_{t_*}]$ and $C_{t_*} = [\tilde{C}_{t_*}, \tau\alpha_{t_*}]$ and replacing α_{t_*} by 0 for all other slices. The optimal step-size τ can again be found by solving (5.50) with $E_{t_*} = \tilde{\Gamma}\alpha_{t_*}\psi_{t_*}^\top$.

On the other hand, if (5.55) is satisfied, we must then check the analogous criteria for Ψ with $\psi = \tilde{\Psi}\beta$:

$$\max_{1\leq t\leq T} \sigma_{max}(W_t\tilde{\Psi}) \leq 1 \tag{5.56}$$

Now, if (5.56) is violated this means we do not need to update Ψ and just update $\Gamma = [\tilde{\Gamma}, \gamma_{t_*}]$ and $C_{t_*} = [\tilde{C}_{t_*}; \tau\beta_{t_*}]$ with β_{t_*} replaced by 0 for all other slices. The optimal step size τ is found by (5.50) with $E_{t_*} = \gamma_{t_*}\beta_{t_*}^\top\tilde{\Psi}^\top$. If this too is satisfied, then we must check the original criteria (2) to potentially update both dictionaries if violated. The order of these global checks can depend on knowledge of the intended sizes of each dictionary. For our purposes we propose to check which of the two violates their constraint more, giving a larger global step. Because (5.55) and (5.56) are lower bounds of (2), satisfying them will not be sufficient to guarantee that we have reached a global minimum and so (2) is still necessary to check in this case. The full algorithm for checking global optimality is outlined in Algorithm 14.

5.5 Experiment: Patch-based Dictionary Learning for dMRI Denoising

For our application we learn our dictionaries from HARDI data though our methods can be applied to any dMRI protocol. Specifically, we experimented on a phantom and a real HARDI brain dataset. The phantom is taken from the ISBI 2013 HARDI Reconstruction Challenge used throughout the thesis, a $V = 50 \times 50 \times 50$ volume consisting of 20 phantom fibers crossing intricately within an inscribed sphere, measured with $G = 64$ diffusion measurements. Our initial experiments test on a 2D 50×50 slice of this data for simplification.

The phantom dataset includes two noise levels: a low noise level of SNR=30 dB and a high noise level of SNR=10 dB. The denoising task will be to denoise the SNR=10 dB data using dictionaries learned from the SNR=30 dB data and record the error with respect to the "ground truth" SNR=30 dB data by calculating Peak SNR (PSNR):

$$PSNR = 10 \log_{10} \frac{MAX_I}{MSE^2}, \tag{5.57}$$

where MAX_I indicates the maximum value in the original SNR=30 dB signal, and MSE is the mean squared error between the original SNR=30 dB signal and the reconstruction. The higher the PSNR, the more accurate the reconstruction will be. We chose a subset of slices of the SNR=30 dB to learn our 2D spatial-angular

dictionaries and used a selection of the remaining slices as test data for denoising.

After validation on phantom data, we show qualitative denoising results on a real HARDI volume with $G = 127$ diffusion measurements using our proposed spatial-angular dictionaries learned on a subset of 2D slices.

5.5.1 Patch-Based Training for dMRI

In theory, our spatial-angular dictionary learning method is capable of learning global spatial and angular dictionaries, $\Psi \in \mathbb{R}^{G \times r_1}$ and $\Gamma \in \mathbb{R}^{V \times r_2}$, over an entire dMRI dataset of size $G \times V$. However, the typical size of a HARDI brain volume is on the order of $V = 100^3$ voxels, and $G = 100$ diffusion measurements, i.e. of size $G \times V = 10^8$. Furthermore, the number of training examples T depends on the size of the training sample. This would require a very large number of training examples of entire dMRI datasets, which is largely infeasible for our algorithm. Because the spatial domain is orders of magnitude larger than the angular domain, one way to curb the computational burden is to reduce our dictionary learning to local spatial patches for all diffusion measurements. Patch-based methods are popular for image processing tasks such as denoising, filtering, inpainting, and object detection [160]. In addition, local dictionaries are beneficial for capturing local features that are often repeated in an image, such as edges, textures or objects.

For training we thus choose a random selection of spatial patches that is consistent along the diffusion domain. For computational simplicity and purposes of visualiza-

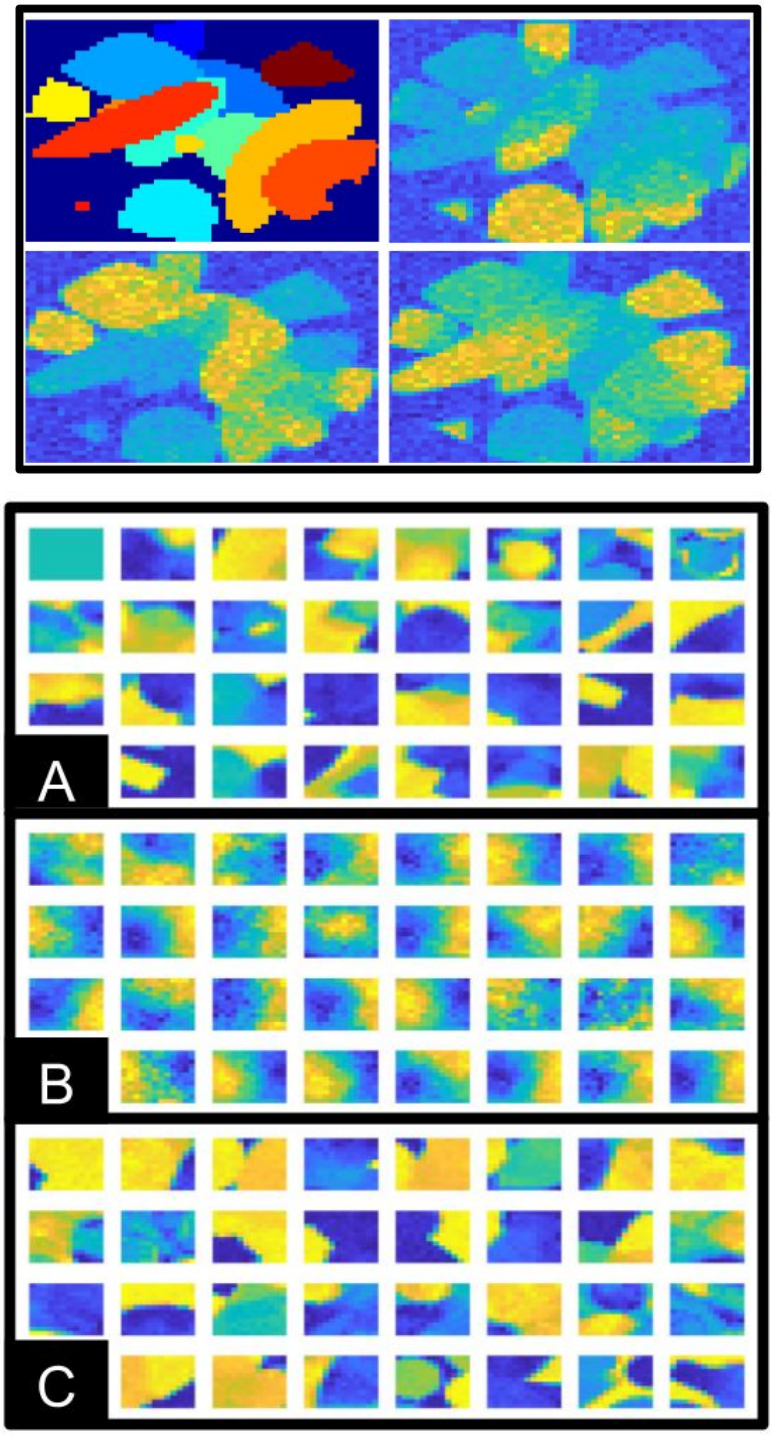


Figure 5.2: Top: Phantom HARDI ground truth fiber segmentations and three diffusion weighted images used for training on patches of size 12×12. Bottom: Spatial patch dictionaries learned via A. KSVD independently from angular dictionary, B. KDRSDL jointly with angular dictionary, C. the proposed method jointly with angular dictionary. B. appears to have reached a local minima farther while A. and C. closely resemble each other and pick up sharp edges and shapes correlated with the training phantom.

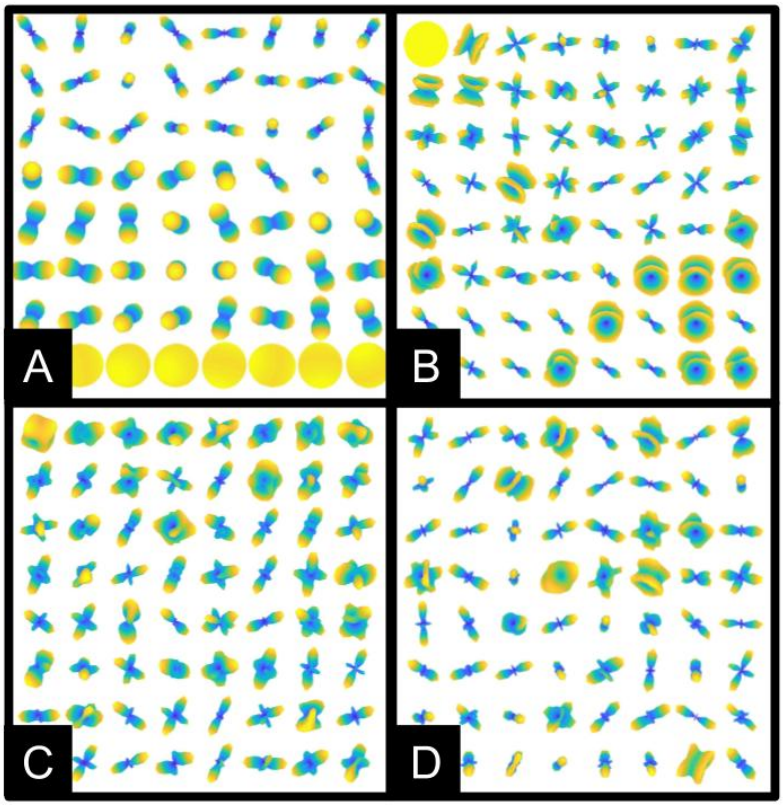


Figure 5.3: Comparison of angular dictionaries. *A*. Fixed spherical ridglets. *B*. – *D*. Angular dictionaries trained on the phantom HARDI data learned via B. KSVD independently from spatial dictionary, *C*. KDRSDL jointly with spatial dictionary, and *D*. the proposed method jointly with spatial dictionary. KSVD and the proposed method produce clean single fiber ODFs while KDRSDL ODFs are noisier.

tion, we limit our initial experiments to 2D spatial patches of size $\sqrt{P} \times \sqrt{P}$ instead of 3D, i.e. $S_t \in \mathbb{R}^{G \times P}$. Depending on the detail and size of an image, popular patch sizes range from $\sqrt{P} = 5$ to 15. For our data, $\sqrt{P} = 12$ gives a good amount of detail and is not too large to process.

5.5.2 Dictionary Learning Comparisons

We will validate our separable dictionary learning method by showing its performance on denoising HARDI data. While there are numerous denoising methodologies in the literature, we will focus on utilizing our learned dictionaries within sparse coding which has been used frequently in the dMRI literature [108]. With our learned spatial and angular dictionaries we use spatial-angular sparse coding, proposed in [84, 85] which solves (5.2) only for C with $T = 1$. For spatial patch-based dictionaries, we will apply sparse coding for each patch and average the results across overlapping patches.

To validate the results of our proposed separable dictionary learning method we consider four dictionary comparisons with respect to the performance of denoising:

1. ***Angular vs. Spatial-Angular:*** *will the proposed spatial-angular framework for dictionary learning and sparse coding outperform state-of-the-art framework for angular dictionary learning and sparse coding for denoising?*

2. ***Fixed vs. Learned:*** *will dictionaries learned from dMRI data outperform fixed analytic dictionaries for denoising?*

3. ***Separate vs. Joint:*** *will learning spatial and angular dictionaries jointly via separable dictionary learning better represent dMRI data than learning spatial and angular dictionaries independently each by classical methods like KSVD?*

4. ***Local vs. Global:*** *will our globally optimal separable dictionary learning out-*

perform other locally optimal separable dictionary learning methods?

For comparison 1, we will compare against state-of-the-art angular dictionary learning and sparse coding frameworks. In particular, we will solve the angular dictionary learning problem (5.1) with the commonly used KSVD algorithm [73]. As an additional comparison to state-of-the-art denoising methods, we will also add spatial regularization based on total-variation (TV) to angular sparse coding [161].

For comparison 2, we will compare against fixed spatial and angular dictionaries used in the dMRI literature: For the angular domain we will use the spherical ridgelet (SR) dictionary popularly used in angular sparse coding and compressed sensing for dMRI [15, 44, 45, 118] (see Figure 5.3 A for visualization). For the spatial domain, we will use curvelets, a popular choices for sparsely representing image data, and a good choice for representing dMRI images as seen in our previous work [84, 85].

For comparison 3, we will use KSVD [73] to learning spatial and angular dictionaries independently. In this regime, the spatial and angular dictionaries have no knowledge of each other. Identifying if our proposed joint learning method is advantageous over the fast and easy KSVD applied to each domain separately is an important question to ask.

Finally, for comparison 4, we utilize a framework called Kronecker-Decomposable Robust Sparse Dictionary Learning (KDRSDL) [157] which is a separable dictionary learning method that does not guarantee globally optimal solutions. KDRSDL solves a low-rank variation of (5.2) which the authors show is useful for background sub-

Comparison	1		2		3		4	
Method	Angular	Spatial-Angular	Fixed	Learned	Separate	Joint	Local	Global
I-SR	✓		✓					
I-SR + TV	✓		✓					
Curve-SR		✓	✓					
I-KSVD	✓			✓	✓		✓	
KSVD-KSVD		✓		✓	✓		✓	
KDRSDL		✓		✓		✓	✓	
Proposed		✓		✓		✓		✓

Table 5.1: Checklist of properties for each dictionary type to compare each method. Purple indicates fixed dictionaries, pink indicates spatial and/or angular dictionaries learned independently, and green indicates a joint spatial-angular dictionary.

		Angular			
		SR	KSVD	KDRSDL	Proposed
Spatial	I	I-SR (+ TV)	I-KSVD		
	Curve	Curve-SR	Curve-KSVD		
	KSVD	KSVD-SR	KSVD-KSVD		
	KDRSDL			KDRSDL	
	Proposed				Proposed

Table 5.2: Organization of spatial and angular dictionaries. Purple indicates fixed dictionaries, pink indicates spatial and/or angular dictionaries learned independently, and green indicates a joint spatial-angular dictionary.

tracting and image denoising.

We use a "Spatial-Angular" notation to keep track of the different dictionary choices, where, for example, I-SR uses the identity for the spatial dictionary and spherical ridgelets for the angular dictionary, I-KSVD learns the angular dictionary using KSVD, and KSVD-KVSD uses the spatial and angular dictionaries learned by KSVD independently. See Table 5.1 for a checklist of the different dictionary properties for each of the 4 comparisons and Table 5.2 for a summary of the spaital and angular domains for each method.

5.5.3 Visualization

In Figures 5.2 and 5.3 we visualize the spatial and angular dictionaries learned from each method on phantom HARDI data as well as the spherical ridgelet dictionary atoms in Figure 5.3 *A*. The learned dictionary atoms are organized left to right from top to bottom by the number of training examples that used each atom, i.e. the number of nonzero coefficients associated to each atom in training. For KSVD, this ordering is independent for the spatial and angular dictionaries, while the atoms resulting from KDRSDL and the proposed method are ordered jointly (without repeats), i.e. the top left spatial and angular atoms combine to create the most utilized spatial-angular dictionary.

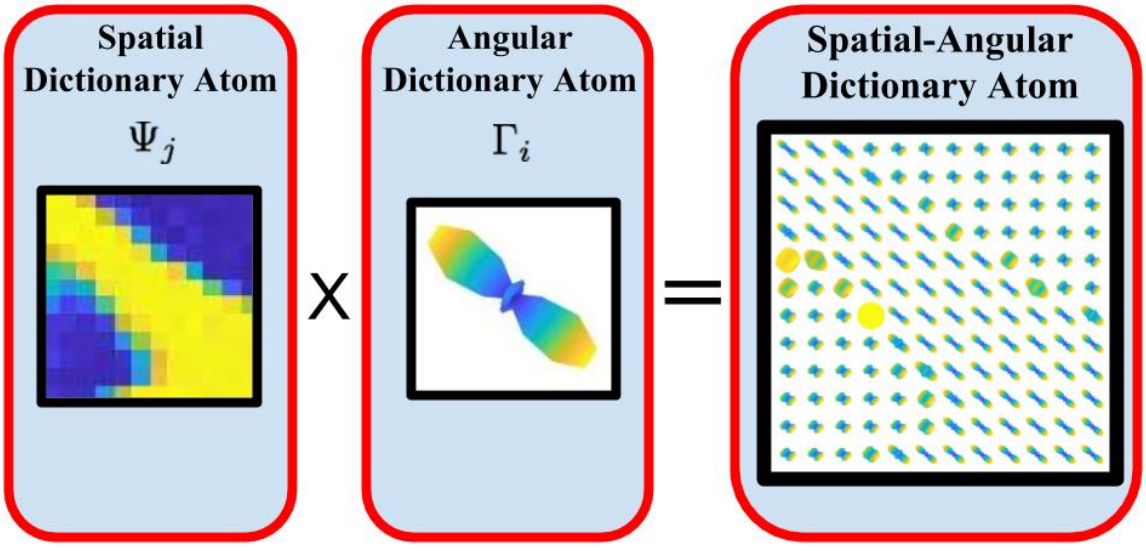


Figure 5.4: Spatial-Angular dictionary atom example learned jointly from phantom HARDI data with the proposed method. We can see that we have the ability to model fiber tracts with very few atoms.

For the spatial dictionaries in Figure 5.2, we notice clear similarities between our

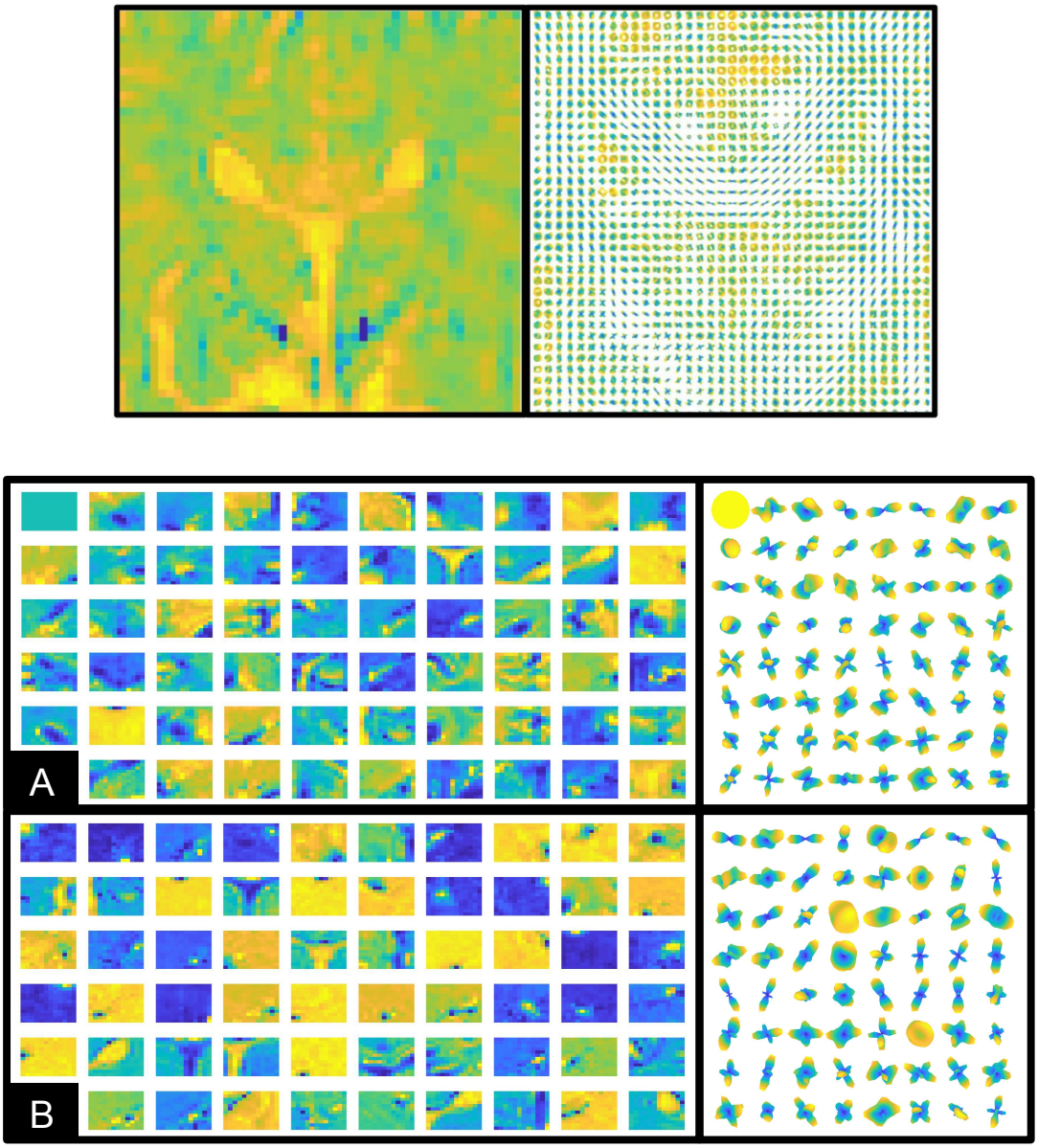


Figure 5.5: Top: Example of real HARDI brain training data, one of the spatial DWIs (top left) and the corresponding ODFs (top right). Bottom: *A*. Spatial and angular dictionaries learned independently via KSVD. Each are sorted (left to right, top to bottom) by their individual frequencies of use in modeling the training data. *B*. Spatial and angular dictionaries learned jointly by the proposed method. Each are sorted (left to right, top to bottom), by their joint frequencies. For example, the top left spatial and angular atoms are together the most frequently used joint spatial-angular atom.

method and the atoms produced by KSVD. In contrast, the spatial atoms produced by KDRSDL are fuzzier, lacking the clearly defined edges and geometric shapes that are evident in the phantom dataset. These shapes resemble atoms that have landed in a local minimum or saddle point, farther from the global minimum reached by our method. This trend is similar for the angular atoms in Figure 5.3. We can see that the results of the proposed method has greater variation in the orientations of single fiber ODFs. The most utilized atoms in KSVD are the purely isotropic atom and the noisy isotropic atoms, whereas the atoms most frequently used with the other methods are the single fiber atoms. In Figure 5.4 we show an example of a single spatial-angular atom learned jointly by our proposed separable dictionary learning method the resembles a fiber tract structure.

Finally, in Figure 5.5 we show spatial and angular dictionaries (bottom) learned from real HARDI brain data (top) for KSVD ($A.$) and our proposed method ($B.$). We notice large structures in the spatial atoms like the CSF region as well as atoms with specific spatial localization that resemble fiber structure. Each atom is sorted (left to right, top to bottom) by their joint frequency of being used to represent the training data. For example, the top left spatial and angular atoms are together the most frequently used joint spatial-angular atom in training.

5.5.4 HARDI Denoising Results

The results of the denoising experiment on the phantom HARDI data are recorded in Table 5.3. We repeated the experiment on three slices of the phantom HARDI data that were not used for training. For each experiment, our reconstruction using the dictionaries learned jointly from our method achieved the highest PSNR values (right-most column of Table 5.3), outperforming both KDRSDL which learns dictionaries jointly but may suffer from local solutions, and KSVD-KSVD which learns spatial and angular dictionaries separately, as well as fixed spatial-angular dictionaries. These results provide a preliminary validation of the importance of separable dictionary learning with global optimality over methods that may converge at a local minimum or saddle point.

Figure 5.6 shows the qualitative results of our denoising experiment in comparison to the denoising results of the SR fixed dictionary with a close-up in Figure 5.7. Then, in Figure 5.8 we show denoising results on real HARDI data using our proposed dictionaries with noticeable regions of improvement highlighted in red. While we have validated our dictionary learning algorithm for the task of denoising, we do not intend to compare against an exhaustive list of denoising methods.

Domain	Angular			Spatial-Angular			
Type	Fixed	Fixed	Separate	Fixed	Separate	Joint	Joint
Method	I-SR	I-SR+TV	I-KSVD	Curve-SR	KSVD-KSVD	KDRSDL	**Proposed**
Slice 25	16.631	16.634	18.011	17.000	19.182	18.793	**19.501**
Slice 30	16.715	16.720	16.090	17.087	17.001	16.725	**17.221**
Slice 35	17.311	17.323	16.679	17.793	17.675	17.418	**17.868**

Table 5.3: Peak Signal-to-Noise Ratio (PSNR) denoising results on three different 2D HARDI phantom image slices. We compared the domains of angular vs spatial-angular sparse coding with dictionaries that are either of type fixed (purple), learned in the spatial and angular domains separately (pink), or learned in the spatial-angular domain jointly (green). Denoising using our proposed joint spatial-angular dictionary learning method with global optimality outperforms denoising with both fixed and learned dictionaries from other methods.

5.6 Conclusion

In this Chapter, we have developed a novel separable dictionary learning method that, to the best of our knowledge, provides the first guarantees of global optimality for this problem. To achieve this, we have framed this problem as a tensor factorization, extending theoretical results from two-factor matrix factorization to the more complex case of three-factor tensor factorization observed in separable dictionary learning.

Given this theoretical justification, we have proposed a novel algorithm to find global minima of the separable dictionary learning problem by alternating between a period of local descent to a stationary point and a check for global optimality. If the global criteria is not satisfied, the algorithm will append an additional dictionary atom and continue the descent to another stationary point. In this way, our algorithm provides a "rank-aware" methodology that could provide low-rank or overcomplete solutions, a reasonable midpoint between the low-rank solutions of KDRSDL and the overcomplete solutions of KSVD. This too depends on the initial dictionary size

which may be application specific. Furthermore, the alteration of updates between each separate dictionary is flexible in our algorithm, and can be tailored to specific *a priori* knowledge of the relative dictionary sizes based on the data.

As a proof of concept, we applied the proposed algorithm to the domain of dMRI which is well suited for our framework due to the separable spatial-angular structure of the data. While most dictionary learning methods for dMRI restrict learning to the angular domain, we learn both spatial and angular dictionaries jointly in this work. We showed in a denoising task that sparse coding using spatial and angular dictionaries learned jointly, outperforms state-of-the-art dMRI denoising algorithms that use sparse coding with angular dictionaries alone. Furthermore, we validated that joint learning provides better reconstructions than the alternative of learning spatial and angular dictionaries independently by simpler methods such as KSVD. Finally, our results indicate that having a globally optimal solution also outperforms methods that arrive at stationary points.

In the next chapter, we will incorporate the patched-based dictionaries learned here within a convolutional sparse coding methodology to relate the local dictionaries to a global image.

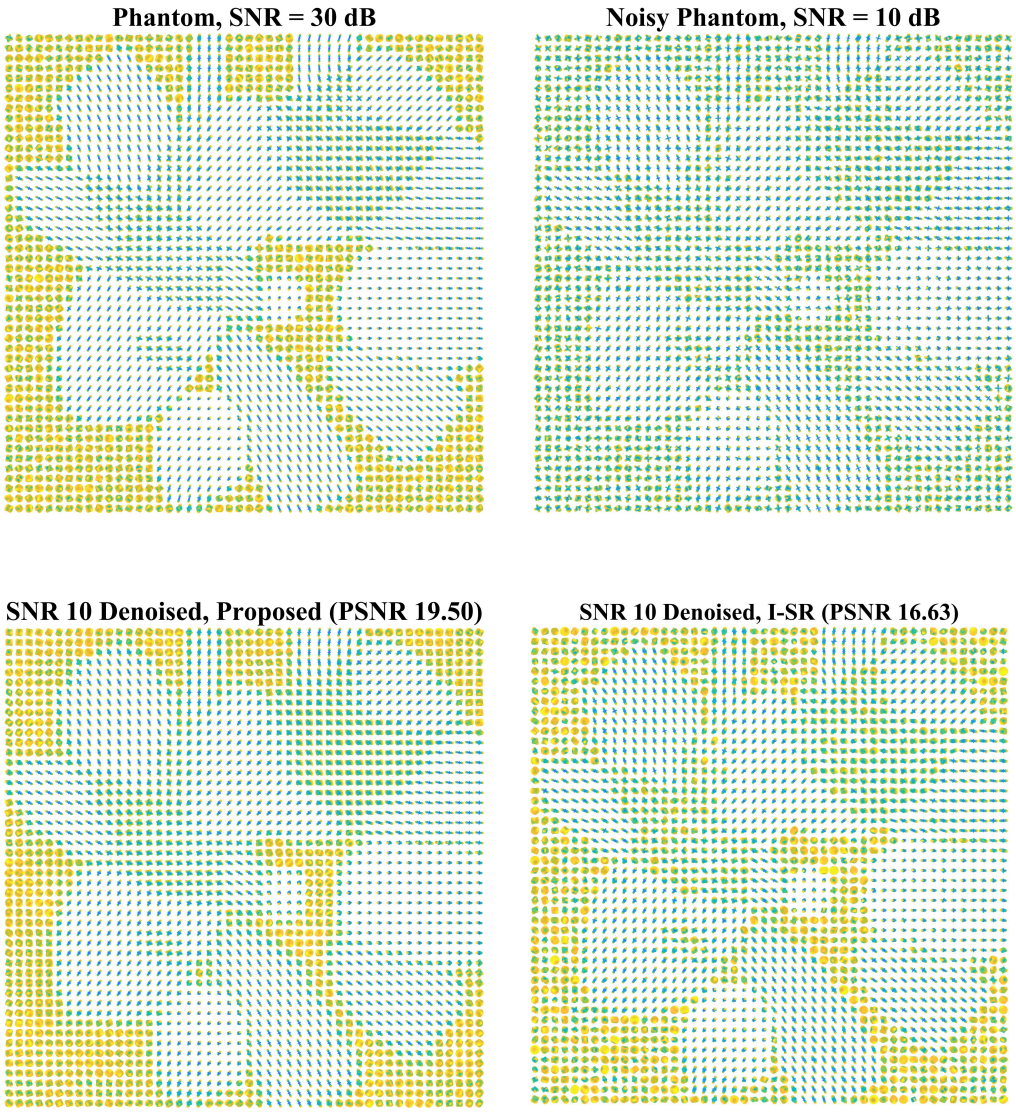


Figure 5.6: Results of HARDI phantom denoising experiment. Top left: Original Phantom data with SNR=30 dB. Top right: Noisy version with SNR=10 dB. Bottom left: Denoised reconstruction of noisy phantom using our learned spatial-angular dictionaries with spatial-angular sparse coding. Bottom right: Denoised reconstruction of noisy phantom using a fixed spherical ridgelet dictionary with angular sparse coding (I-SR). We notice our proposed method produces a more accurate reconstruction in comparison to the original SNR=30 dB. For more detailed visualization see the close-ups in Figure 5.7.

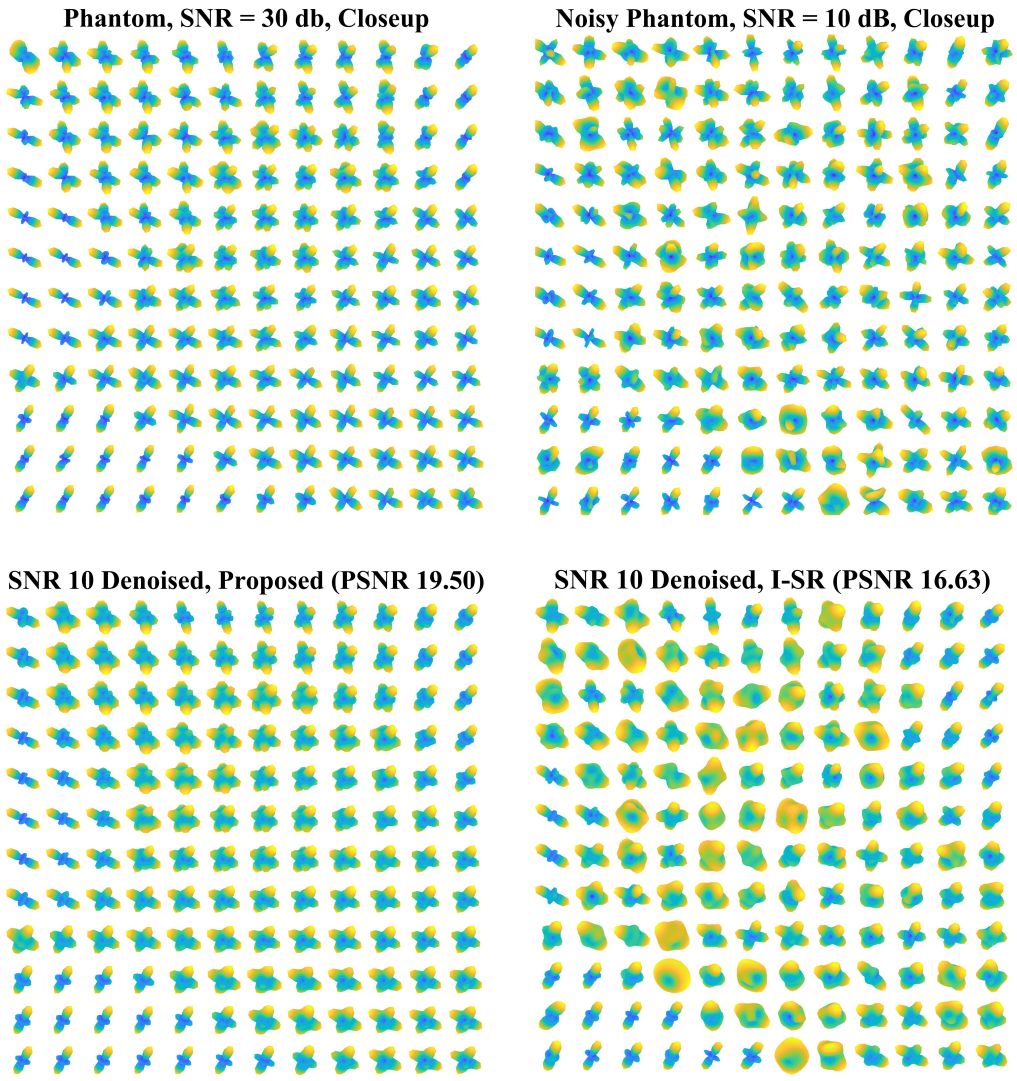


Figure 5.7: Close-ups of HARDI phantom denoising results from Figure 5.6. The reconstruction of the noisy SNR=10 dB HARDI phantom (top right) using our proposed spatial-angular dictionary (bottom left) produces a more accurate denoised reconstruction in comparison to the original phantom with SNR=30 dB (top left), than for the fixed spherical ridgelet (SR) dictionary (bottom right).

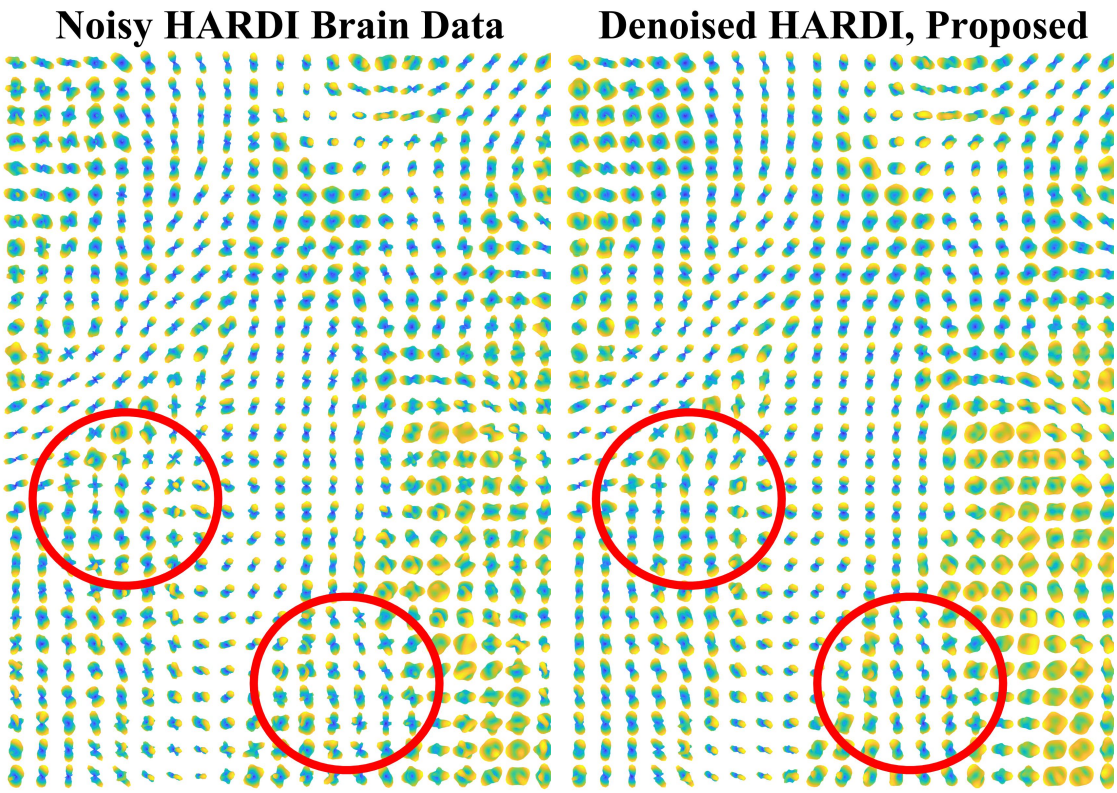


Figure 5.8: Denoising Real HARDI brain data. Left: Original noisy HARDI brain region. Right: Denoised reconstruction using our learned spatial-angular dictionaries within our spatial-angular sparse coding.

Chapter 6

Future Directions: Convolutional Spatial-Angular Sparse Coding

6.1 Introduction

In Chapter 5, we learned spatial-angular dictionaries directly from dMRI data with the aim to discover more compact representations for sparse coding and ultimately compressed sensing. But, because of the size of dMRI and the need for multiple training examples during learning, we were restricted to learning local patch-based dictionaries in the spatial domain. While it is possible to sparsely reconstruct each patch, the global sparsity level will be restricted by the number of patches in the image and undesirable artifacts can occur around the boundaries of the different patches. Therefore, to make comparisons with the spatial-angular sparse coding results of

Chapter 3 that used fixed, global dictionaries, we need a mechanism to relate the local sparse reconstruction of patch-based dictionaries with the global signal. Introduced in the background Section 2.2.4, this mechanism is convolution and is used within a framework known as convolutional sparse coding.

In this chapter, we extend the methods of convolutional sparse coding of images to the case of spatial-angular convolutional sparse coding for dMRI, or more generally, convolutional sparse coding with separable dictionaries. We propose a formulation of the convolutional sparse coding problem for separable dictionaries that relies on tensor decomposition to handle large-scale dMRI data. We provide a preliminary result comparing the global reconstruction of convolutional spatial-angular sparse coding with that of the traditional spatial-angular sparse coding in Chapter 3 with fixed dictionaries.

As an important note, this chapter remains an initial formulation for future efforts, and serves as a starting point for a number of possible extensions and applications yet to be fully developed. We henceforth present only the groundwork ideas for future directions in convolutional spatial-angular sparse coding.

6.2 Problem Formulation

Following the separable spatial-angular setting adopted throughout this thesis, we can write a dMRI signal $S \in \mathbb{R}^{G \times V}$ as $S = \Gamma C \Psi^\top$, where $\Gamma \in \mathbb{R}^{G \times N_\Gamma}$ is the angular

dictionary, $\Psi \in \mathbb{R}^{V \times N_\Psi}$ is the spatial dictionary and $C \in \mathbb{R}^{N_\Gamma \times N_\Psi}$ are the spatial-angular coefficients. In this chapter, we assume that a patch-based dictionary for the spatial domain is given as $D \in \mathbb{R}^{P \times N_D}$, with N_D atoms $D_k \in \mathbb{R}^P$, where $P << V$, such as the dictionary learned from Chapter 5. Then, to relate D to a global spatial dictionary Ψ, we can build the atoms of Ψ as shifted versions of the atoms of D in the spatial domain.

To see this, each atom D_k, of the patch-based dictionary D, can be viewed as a spatial filter of size $\sqrt{P} \times \sqrt{P}$ (written here as 2D for simplicity and to be consistent with the 2D patch-based dictionaries learned in Chapter 5). Likewise, the atoms Ψ_j of global spatial dictionary Ψ, can be viewed as images with V voxels. Then, atom Ψ_1, for instance, will be the filter D_1 centered at voxel location $(1,1)$[1] and zero elsewhere in the image. Then, atom Ψ_2 will be the same filter D_1 shifted to voxel location $(1,2)$ and zero elsewhere. Once D_1 has been shifted for all V voxels in the image, we repeat the process for D_2, such that Ψ_{V+1} is the filter D_2 located at voxel location $(1,1)$ with zeros elsewhere and so forth. Each Ψ_j image is then vectorized to form the columns of Ψ.

Mathematically, each global spatial atom Ψ_j can be written as a convolution of patch-dictionary D_k with the Dirac delta map δ_l which has a 1 at voxel location l

[1]We assume appropriate boundary conditions like zero padding or circularity to accommodate the size of the filter.

and zero else where in the image, i.e. for all voxel v

$$\Psi_j(v) = (D_k * \delta_l)(v), \tag{6.1}$$

where $*$ is a 2D convolution (see background Section 2.2.4 for a definition) and $j = (k, l)$. Here the convolution with δ_l acts as a simple placement operator, placing filter D_k at location l. In fact, this construction allows for variable voxel locations in the image, for example, strides of $l = 1, 3, 5, \ldots$ or $l = 5, 10, 15, \ldots$, or even arbitrary locations, where $L \leq V$ is the total number of voxel locations and $N_\Psi = N_D L$.

Now, to write a dMRI signal in terms of shifted versions of patch-based dictionaries, we will build from our separable spatial-angular model as seen in (3.12):

$$\begin{aligned}
\mathcal{S}(v, q) &= \sum_{i=1}^{N_\Gamma} \sum_{j=1}^{N_\Psi} c_{i,j} \Gamma_i(q) \Psi_j(v) && (6.2)\\
&= \sum_{i=1}^{N_\Gamma} \sum_{l=1}^{L} \sum_{k=1}^{N_D} c_{i,k,l} \Gamma_i(q) (D_k * \delta_l)(v) && (6.3)\\
&= \sum_{l=1}^{L} \left(\left[\sum_{i=1}^{N_\Gamma} \sum_{k=1}^{N_D} c_{i,k,l} \Gamma_i(q) D_k \right] * \delta_l \right)(v). && (6.4)
\end{aligned}$$

Here, we have expanded the coefficient matrix $C = [c_{i,j}] \in \mathbb{R}^{N_\Gamma \times N_\Psi}$ as a three dimensional tensor $\underline{C} = [c_{i,k,l}] \in \mathbb{R}^{N_\Gamma \times N_D \times L}$. Using our tensor notation from Chapter 5, we let $C_l \in \mathbb{R}^{N_\Gamma \times N_D}$ be the l^{th} slice of $\underline{C}$. Then, with $S \in \mathbb{R}^{G \times V}$, (6.4) becomes:

$$S = \sum_{l=1}^{L} \left(\sum_{i=1}^{N_\Gamma} \sum_{k=1}^{N_D} c_{i,k,l} \Gamma_i D_k \right) * \delta_l = \sum_{l=1}^{L} \Gamma C_l D^\top * \delta_l, \tag{6.5}$$

where, with abuse of notation, the convolution on the right hand side of (6.5) acts over each of the G rows of $\Gamma C_l D^\top \in \mathbb{R}^{G \times P}$. With this representation of the signal S, we can write the convolutional spatial-angular sparse coding problem as:

$$\min_{\{C_l\}} \frac{1}{2} || \sum_{l=1}^{L} (\Gamma C_l D^\top * \delta_l) - S||_F^2 + \lambda \sum_{l=1}^{L} ||C_l||_1. \tag{6.6}$$

One possible way to solve a sparse coding problem is by the Fast Iterative Shrinkage Algorithm (FISTA). Here we propose to apply a convolutional variation of our proposed Kronecker FISTA Algorithm 9 of Chapter 3 by taking special care of the convolution operator when taking the gradient of (6.6). In the next section, we we introduce our proposed algorithm to solve the convolutional spatial-angular sparse coding problem.

As an important remark, our proposed convolutional sparse coding formulation in (6.6) does not, a priori, exactly match the more standard form seen in Section 2.2.4 which involve convolutions of feature maps x_k with the local patch dictionaries D_k. However, if we assume $L = V$ (i.e. when each voxel location is considered for patch placement), the formulation we propose may be written in a form that exactly extends the one of (2.57) to the case of spatial-angular signals. Indeed, going back to (6.4)

and rearranging the sum, we have:

$$\begin{aligned}
\mathcal{S}(v,q) &= \sum_{i=1}^{N_\Gamma}\sum_{l=1}^{V}\sum_{k=1}^{N_D} c_{i,k,l}\Gamma_i(q)(D_k * \delta_l)(v) \\
&= \sum_{k=1}^{N_D}\sum_{i=1}^{N_\Gamma}\Gamma_i(q)\sum_{l=1}^{V} c_{i,k,l}(D_k * \delta_l)(v) \\
&= \sum_{k=1}^{N_D}\sum_{i=1}^{N_\Gamma}\Gamma_i(q)\sum_{l=1}^{V} c_{i,k,l}\sum_{p=1}^{P} D_k(p)\delta_l(v-p) \\
&= \sum_{k=1}^{N_D}\sum_{i=1}^{N_\Gamma}\Gamma_i(q)\sum_{p=1}^{P} D_k(p)\sum_{l=1}^{V} c_{i,k,l}\delta_l(v-p) \\
&= \sum_{k=1}^{N_D}\sum_{i=1}^{N_\Gamma}\Gamma_i(q)\sum_{p=1}^{P} D_k(p)c_{i,k,v-p} \\
&= \sum_{k=1}^{N_D}\sum_{i=1}^{N_\Gamma}\Gamma_i(q)(D_k * \underline{C}_{i,k,\cdot})(v)
\end{aligned}$$

or in matrix form

$$S = \sum_{k=1}^{N_D}\Gamma(D_k * X_k)$$

where $X_k = \underline{C}_{\cdot,k,\cdot} \in \mathbb{R}^{N_\Gamma \times V}$ is the k-th slice of the coefficient tensor with respect to the second coordinate and $D_k * X_k \in \mathbb{R}^{N_\Gamma \times V}$ is the 2D convolution of the patch dictionary D_k applied to each row of X_k. Then, the convolutional sparse coding problem (6.6) becomes equivalent to:

$$\min_{\{X_k\}} \frac{1}{2}||\sum_{k=1}^{N_D}\Gamma(D_k * X_k) - S||_F^2 + \lambda\sum_{k=1}^{N_D}||X_k||_1, \tag{6.7}$$

and, extending the framework of Section 2.2.4, X_k can be here interpreted as a feature

map of coefficients of the signal in the angular dictionary. Possible methods to solve (6.7) could include extensions of algorithms used to solve the classical (2.57) such as convolutional FISTA [83] or to convert the convolution operation to an element-wise multiplication using the Fourier Transform and applying ADMM [74–76, 78]. Exploring these alternative methods will be the focus of future work as well as their comparison in terms of computational efficiency with the proposed algorithm of the next section.

6.3 Algorithm

To solve (6.6) using FISTA, we need the gradient with respect to each C_l. With this in mind, we can define an operator to abstract the summation of convolutions. First let $\mathcal{P}_l := \Gamma C_l D^\top$. Then, using again the tensor notation from Chapter 5, $\mathcal{P}_l$ can be viewed as the l^{th} slice of tensor $\underline{\mathcal{P}} := \underline{C} \times_1 \Gamma \times_2 D \in \mathbb{R}^{G\times P\times L}$, where $\times_n$ denotes the n-mode product of a tensor by a matrix. Then we define the operator $\mathcal{L}$ as:

$$\mathcal{L} : \mathbb{R}^{G\times P\times L} \longrightarrow \mathbb{R}^{G\times V}, \quad \underline{\mathcal{P}} \longmapsto S = \sum_{l=1}^{L} \mathcal{P}_l * \delta_l. \tag{6.8}$$

This construction allows us to define the adjoint of $\mathcal{L}$ as:

$$\mathcal{L}^* : \mathbb{R}^{G\times V} \longrightarrow \mathbb{R}^{G\times P\times L}, \quad S \longmapsto \underline{\mathcal{P}}, \quad \underline{\mathcal{P}}_{q,p,l} = S_{q,p+l}, \tag{6.9}$$

which takes the $\sqrt{P} \times \sqrt{P}$ patch of S centered at voxel location l, for all G gradient directions, vectorizes them to row vectors of length P and creates the matrix $\mathcal{P}_l \in \mathbb{R}^{G \times P}$. (Voxels at boundary locations can be accounted for by zero padding the image or by other common heuristics in image processing). This is then repeated for all L voxel locations, concatenating each $\mathcal{P}_l$ slice in the third dimension as tensor $\underline{\mathcal{P}} \in \mathbb{R}^{G \times P \times L}$. Then, the convolutional spatial-angular sparse coding problem (6.6) can be rewritten as:

$$\min_{\underline{C}} \frac{1}{2}||\mathcal{L}(\underline{C} \times_1 \Gamma \times_2 D) - S||_F^2 + \lambda||\underline{C}||_1 \tag{6.10}$$

Then the gradient of $h := \frac{1}{2}||\mathcal{L}(\underline{C} \times_1 \Gamma \times_2 D) - S||_F^2$ with respect to $\underline{C}$ is :

$$\nabla_{\underline{C}} h = \mathcal{L}^*(\mathcal{L}(\underline{C} \times_1 \Gamma \times_2 D)) \times_1 \Gamma^\top \times_2 D^\top - \mathcal{L}^*(S) \times_1 \Gamma^\top \times_2 D^\top. \tag{6.11}$$

For efficiency, $\hat{S} := \mathcal{L}^*(S) \times_1 \Gamma^\top \times_2 D^\top$ is pre-computed. In addition, since $\mathcal{L}$ does not act on the angular domain, we can pass the multiplication $\times_1 \Gamma^\top$ inside the operator and pre-compute $\Gamma^\top \Gamma$ to get:

$$\nabla_{\underline{C}} h = \mathcal{L}^*(\mathcal{L}(\underline{C} \times_1 \Gamma^\top \Gamma \times_2 D)) \times_2 D^\top - \hat{S} \tag{6.12}$$

The proposed algorithm is presented in Algorithm 15 and called Convolutional Tensor FISTA (Conv-Ten-FISTA). The bulk of computations in (6.12) lies in the tensor

multiplications which lead to a complexity of the order of $O(LN_D(N_\Gamma^2 + N_\Gamma P + GP))$.

Algorithm 15 Convolutional Tensor FISTA (Conv-Ten-FISTA)

Given: Γ, D.
Initialize: $\underline{Z}_0 = \underline{C}_0 = 0, n_1 = 1, \ell, \lambda, \epsilon$.
Precompute: $\hat{S} = \mathcal{L}^*(S) \times_1 \Gamma^\top \times_2 D^\top$.
while error $> \epsilon$ **do**
  $\ell = \text{linesearch}(\underline{Z}_k)$
  $\nabla_{\underline{Z}_k} h = \mathcal{L}^*(\mathcal{L}(\underline{Z}_k \times_1 \Gamma^\top \Gamma \times_2 D)) \times_2 D^\top - \hat{S}$
  $\underline{C}_k = \text{shrink}_{\lambda/\ell}(\underline{Z}_k - \lambda \nabla_{\underline{Z}_k} h / \ell)$
  $n_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_k^2})$
  $\underline{Z}_{k+1} = \underline{C}_k + \frac{n_k - 1}{n_{k+1}}(\underline{C}_k - \underline{C}_{k-1})$
  $k \to k + 1$
end while

6.4 Preliminary Results

In this section, we provide a preliminary result of our proposed convolutional spatial-angular sparse coding method using patch-based spatial-angular dictionaries learned previously in Chapter 5. Using the 12×12 spatial and angular dictionaries learned by our proposed separable dictionary learning method, K-SVD [73] and the recent KDRSDL [157] we reconstruct an entire slice of the phantom HARDI dataset with Conv-Ten-FISTA. These results are then compared against the reconstruction based on the global Curve-SR dictionary and the Kron-FISTA algorithm of Chapter 3. The quantitative results of residual error vs. sparsity are displayed in Figure 6.1.

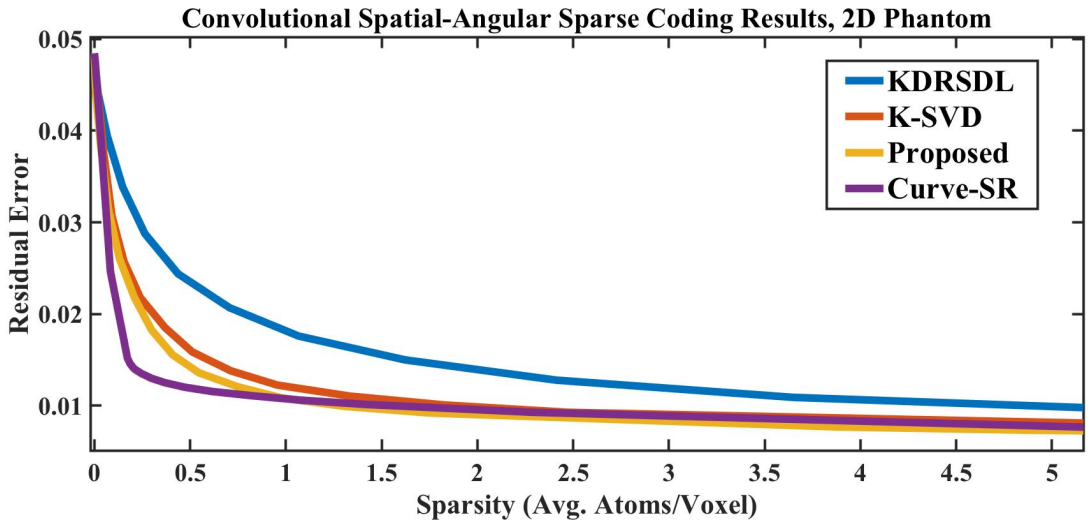


Figure 6.1: Quantitative result of residual error vs. sparsity for the reconstruction of a 50×50 phantom HARDI signal using 12×12 patch spatial-angular dictionaries with the proposed convolutional spatial-angular sparse coding method. We compared the dictionaries learned independently via K-SVD and jointly using KDRSDL and our proposed dictionary learning method in Chapter 5. We also compared against the fixed SR angular dictionary and curvelet spatial dictionary (Curve-SR) using the usual spatial-angular sparse coding without convolution. In this experiment, the Curve-SR outperforms the learned dictionaries. This may be due to the measure of global sparsity that is not representative of the patch-based learned dictionaries and the lack of multiple patch sizes.

The results for that particular example are consistent with the previous observations in the sense that the dictionary learned by the proposed approach of Chapter 5 still performs better than the K-SVD and KDRSDL dictionaries in this convolutional setting. However, at this point, we see that the fixed global dictionary Curve-SR outperforms the learned patch-based dictionaries. A possible explanation for this is that atoms in the curvelet dictionary involve multiple different scales, while our learned dictionaries have a predefined and fixed patch size. In the next section, we formulate a multi-scale extension to the proposed convolutional spatial-angular sparse coding method.

6.5 Multi-Scale Extension

Until now, we have considered convolutional sparse coding for a spatial dictionary of single patch size $\sqrt{P} \times \sqrt{P}$. However, like most analytic dictionaries such as wavelets, spatial filter banks are comprised of dictionary atoms with varying scale or patch size and, depending on the structure of the data, sparse coding may benefit from combining atoms of multiple scales. In this section, we briefly discuss how to incorporate multi-scale dictionaries (each given or learned for example from the previous dictionary learning approach of Chapter 5) within our convolutional sparse coding framework. Unlike previous formulations like [80] which may require special care to use multi-scale dictionaries when considering their global circulant dictionary construction, the addition of multiple dictionary scales in our tensor formulation is quite natural. Let $\mathcal{D} := \{D^m\}_{m=1}^M$ be a set of spatial patch-based dictionaries where $D^m \in \mathbb{R}^{P_m \times N_{D^m}}$ has patch scale P_m and N_{D^m} atoms. As the size of the patches P_m differ, it will be desirable to have various spatial stride rates. Let $\{l_m\}$ be the voxel locations for the dictionary of scale m with a total of L_m locations.

Next, we have the associated set of coefficients $\underline{\mathcal{C}} := \{\underline{C}^m\}_{m=1}^M$ for each scale, where $\underline{C}^m \in \mathbb{R}^{N_\Gamma \times N_{D^m} \times L_m}$ is a tensor and $\underline{\mathcal{P}}^m := \underline{C}^m \times_1 \Gamma \times_2 D^m \in \mathbb{R}^{G \times P_m \times L_m}$. Then, like operator $\mathcal{L}$ defined in (6.13), which takes in the tensor product of coefficients and

dictionaries and produces a signal reconstruction, we can write:

$$\mathcal{L}_m : \mathbb{R}^{G\times P_m\times L_m} \longrightarrow \mathbb{R}^{G\times V}, \quad \underline{\mathcal{P}}^m \longmapsto S^m = \sum_{l_m=1}^{L_m} \mathcal{P}_{l_m}^m * \delta_{l_m}, \tag{6.13}$$

where for each scale m, $\mathcal{P}_{l_m}^m$ is the ${l_m}^{th}$ slice of tensor $\underline{\mathcal{P}}^m$ and $S^m := \mathcal{L}_m(\underline{C}^m \times_1 \Gamma \times_2 D^m) \in \mathbb{R}^{G\times V}$.

Then, to reconstruct the full signal S from each of the S^m we must add such that $S = \sum_{m=1}^{M} S^m$, which we write as the following function $\mathcal{M}$ of the sets of coefficients and dictionaries $\underline{\mathcal{C}}$, Γ, and $\mathcal{D}$:

$$\mathcal{M}(\underline{\mathcal{C}}, \Gamma, \mathcal{D}) := \sum_{m=1}^{M} \mathcal{L}_m(\underline{C}^m \times_1 \Gamma \times_2 D^m) = \sum_{m=1}^{M} S^m = S. \tag{6.14}$$

As $S = \mathcal{M}(\underline{\mathcal{C}}, \Gamma, \mathcal{D})$, we will need the adjoint map $\mathcal{M}^*(S)$ defined as the set of operators $\mathcal{M}^* = \{\mathcal{L}_m^*\}_{m=1}^M$ such that,

$$\mathcal{L}_m^* : \mathbb{R}^{G\times V} \longrightarrow \mathbb{R}^{G\times P_m\times L_m}, \quad S \longmapsto \underline{\mathcal{P}}^m, \tag{6.15}$$

which, as is the same for (6.9), take the $\sqrt{P_m}\times\sqrt{P_m}$ (for 2D filters) patch of S located at voxel location l_m for all G gradient directions, vectorizes them to row vectors of length P_m and creates the matrix $\mathcal{P}_{l_m}^m \in \mathbb{R}^{G\times P_m}$ for each scale m.

With all of these components, our proposed multi-scale convolutional sparse cod-

ing problem becomes

$$\min_{\underline{\mathcal{C}}} \frac{1}{2}||\mathcal{M}(\underline{\mathcal{C}}, \Gamma, \mathcal{D}) - S||_F^2 + \lambda \sum_{m=1}^{M} ||\underline{C}^m||_1. \tag{6.16}$$

To solve this using FISTA, we need to take the gradient of $h = \frac{1}{2}||\mathcal{M}(\underline{\mathcal{C}}, \Gamma, \mathcal{D}) - S||_F^2$ with respect to each $\underline{C}^m$:

$$\nabla_{\underline{C}^m} h = \mathcal{L}_m^*(\mathcal{M}(\underline{\mathcal{C}}, \Gamma, \mathcal{D})) \times_1 \Gamma^\top \times_2 D^{m\top} - \mathcal{L}_m^*(S) \times_1 \Gamma^\top \times_2 D^{m\top}. \tag{6.17}$$

As an extension of Conv-Ten-FISTA to spatial dictionaries with multiple scales, our algorithm Multi-Scale Convolutional Tensor FISTA (MS-Conv-Ten-FISTA) is presented in Algorithm 16. Implementation of the multi-scale convolutional tensor

Algorithm 16 Multi-Scale Convolutional Tensor FISTA (MS-Conv-Ten-FISTA)

Given: $\Gamma, \mathcal{D} = \{D^m\}_{m=1}^M$.
Initialize: $\underline{\mathcal{Z}}_0 = \underline{\mathcal{C}}_0 = \{\underline{C}^m\}_{m=1}^M = \{0\}_{m=1}^M, n_1 = 1, \ell, \lambda, \epsilon$.
Precompute: $\hat{S}^m = \mathcal{L}_m^*(S) \times_1 \Gamma^\top \times_2 D^{m\top}$.
while error $> \epsilon$ **do**
  $\ell = \text{linesearch}(\underline{\mathcal{Z}}_k)$
  $\nabla_{\underline{Z}_k^m} h = \mathcal{L}_m^*(\mathcal{M}(\underline{\mathcal{Z}}_k, \Gamma, \mathcal{D})) \times_1 \Gamma^\top \times_2 D^{m\top} - \hat{S}^m \ \ \forall\ m = 1, \ldots, M$
  $\underline{C}_k^m = \text{shrink}_{\lambda/\ell}(\underline{Z}_k^m - \lambda \nabla_{\underline{Z}_k^m} h/\ell) \ \ \forall\ m = 1, \ldots, M$
  $n_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4n_k^2})$
  $\underline{Z}_{k+1}^m = \underline{C}_k^m + \frac{n_k - 1}{n_{k+1}}(\underline{C}_k^m - \underline{C}_{k-1}^m) \ \ \forall\ m = 1, \ldots, M$
  $k \to k + 1$
end while

FISTA algorithm is left to future work. Incorporating multiple scales of patch-based dictionaries to convolutional sparse coding is one potential method to enhance the accuracy of sparse reconstruction compared to fixed global dictionaries like wavelets

and curvelets used in Chapter 3. Next, as another potential method, we investigate different metrics for sparse regularization which take into account the local sparsity at the patch level.

6.6 Discussion

In this chapter, we have developed novel algorithms connecting patch-based dictionaries like those learned in Chapter 5 to global signal reconstruction of dMRI. We have proposed a convolutional spatial-angular sparse coding method which provides a computationally efficient way to find sparse representations in separable dictionaries for large-scale data like dMRI. Future work will be to further consider the incorporation of multiple scales to validate the performance of patch-based dictionaries over that of global spatial dictionaries like wavelets or curvelets used in Chapter 3.

Another interesting direction to evaluate the comparative results of convolutional sparse coding is the choice of regularizer. Using the global L_0 or L_1 regularizers may in fact be too constraining for convolutional sparse coding because nearby shifted versions of a local dictionary will likely use a redundant set of coefficients, leading to higher than desired global sparsity levels. Furthermore, in the context of compressed sensing, for a global dictionary that comprises shifted versions of a small local patch-based dictionary, coherence will be maximal, and so global sparsity levels may not be useful for guarantees of recovery. In an attempt to improve recovery guarantees

for convolutional sparse coding, the works of [81, 82] propose an interesting notion of local sparsity and expand the ideas of coherence and the restricted isometry property to the patch level, from which they derive new guarantees of global signal recovery for convolutional sparse coding of natural images [79, 80]. These metrics could be easily extended to the present case of convolutional spatial-angular sparse coding, which would be an interesting direction for further improvement of our model.

Finally, as has been discussed throughout this thesis, since the ultimate goal of spatial-angular sparse coding is to accelerate the acquisition of dMRI through (k, q) compressed sensing, future work will be to expand convolutional spatial-angular sparse coding for (k, q) compressed sensing. The work of [162] propose a compressed sensing method using convolutional dictionaries for the application of dynamic MRI. Future efforts can be taken to adapt this model to the spatial-angular setting of dMRI. These areas will be the topic of future work as we consider optimally accelerating dMRI acquisition using learned dictionaries and potentially learning subsampling schemes with a convolutional (k, q) compressed sensing framework.

Chapter 7

Conclusion

For diffusion magnetic resonance imaging, the voxel has been the building block for acquisition, modeling, computation, processing and analysis, from PDF estimation, de-noising, and feature extraction, to higher level tasks like segmentation, registration, classification, and tractography. In this thesis, we have proposed to "think outside the voxel," by utilizing a global view of dMRI that builds from a joint spatial-angular representation of the data instead of the traditional per-voxel angular viewpoint. We have presented three major contributions, and future directions, which illustrate the power of our global representation over local frameworks.

In the first contribution in Chapter 3, we showed the power of joint spatial-angular sparse coding in its potential to drastically decrease the global sparsity of a dMRI representation using a separable spatial-angular dictionary. We showed that our proposed framework can surpass the sparsity limitations imposed by local voxel-wise

sparse coding methods. We additionally developed an array of efficient separable sparse coding algorithms to handle the size of large dMRI data which greatly outperform the state-of-the-art in terms of computation time. Furthermore, this global representation may have additional applications for dMRI processing like de-noising, segmentation, tractography, classification, super-resolution, feature extraction, and compressed sensing.

In our second contribution in Chapter 4, we illustrated the impact of spatial-angular dictionaries on the application to compressed sensing (CS) for dMRI. One key ingredient in CS is the sparsity level of the underlying representation. Under certain theoretical considerations, lowering the sparsity level of a representation can reduce the number of measurements needed to recover a full signal from noisy samples. For dMRI, prior (k, q)-CS methods utilize local sparsity in the spatial and angular domains, which limited the minimum global sparsity level. We have proposed a new (k, q)-CS which exploits the joint spatial-angular sparsity provided by our previous contribution. In direct comparison to these prior methods, we illustrated a great reduction of the number of samples needed for accurate reconstruction and hope this may unlock a new realm of dMRI acceleration after experimenting with optimal sampling schemes and dictionaries.

In our third contribution in Chapter 5, we investigated optimizing our choice of spatial-angular dictionary through dictionary learning. Dictionary learning allows us to estimate dictionaries that are more representative of the data with the potential to

further increase sparsity levels. Following the voxel-wise viewpoint, state-of-art dictionary learning methods for dMRI have restricted themselves to the angular domain with added spatial regularization. To the best of our knowledge, we have proposed for the first time a joint spatial-angular dictionary learning framework to learn separable dictionaries directly from dMRI data. We have posed separable dictionary learning problem as a matrix factorization, which allowed to develop the first guarantees of global optimality for learning separable dictionaries. We then illustrated the superior performance of globally optimal dictionaries compared to alternative dictionary learning methods for the task of dMRI de-noising.

Due to the challenging size of dMRI data, in Chapter 5, we have restricted to learning local spatial patch-based dictionaries instead of dictionaries defined on the entire volumes. To bridge the gap between local patch-based dictionaries and global signal reconstruction, in Chapter 6 we proposed a method of convolutional sparse coding applied to the spatial-angular structure of dMRI. In the absence of proper validation and model comparison, this remains preliminary work as a future direction to further accelerate dMRI acquisition using our learned patch-based dictionaries.

In this thesis, we have shown advanced performance of a joint spatial-angular representation compared to voxel-wise frameworks in the machine learning domains of sparse coding, compressed sensing, and dictionary learning. We hope that our core contributions of a global dMRI representation may be utilized for other dMRI processing and analysis applications like spatial-angular de-noising, spatial-angular fiber

tract segmentation, global feature extraction, global tractography, and global diffusion modeling. Furthermore, we believe this general framework can be extended to incorporate other domains like the 1D temporal diffusion domain which has recently been considered in diffusion modeling or other medical imaging modalities like functional MRI and dynamic MRI which embody similar data structures with separable domains. Our proposed methods and algorithms are general to the case of separable dictionaries and we hope that this thesis will have an impact on the greater signal processing, computer vision, and machine learning communities.

Bibliography

[1] J.-D. Tournier, S. Mori, and A. Leemans, "Diffusion tensor imaging and beyond," *Magnetic Resonance in Medicine*, vol. 65, no. 6, pp. 1532–1556, 2011.

[2] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann *et al.*, "Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection," *The Lancet Neurology*, vol. 14, no. 10, pp. 1037–1053, 2015.

[3] S. Mori, S. Wakana, P. C. Van Zijl, and L. Nagae-Poetscher, *MRI atlas of human white matter*. Elsevier, 2005.

[4] G. Zappala, M. T. de Schotten, and P. J. Eslinger, "Traumatic brain injury and the frontal lobes: what can we gain with diffusion tensor imaging?" *Cortex*, vol. 48, no. 2, pp. 156–165, 2012.

[5] P. Hagmann, M. Kurant, X. Gigandet, P. Thiran, V. J. Wedeen, R. Meuli, and J.-P. Thiran, "Mapping human whole-brain structural networks with diffusion MRI," *PloS one*, vol. 2, no. 7, p. e597, 2007.

BIBLIOGRAPHY

[6] M. Mahesh, "The essential physics of medical imaging, third edition." *Medical Physics*, vol. 40, no. 7, pp. 077 301–n/a, 2013, 077301. [Online]. Available: http://dx.doi.org/10.1118/1.4811156

[7] C. B. Paschal and H. D. Morris, "K-space in the clinic," *Journal of Magnetic Resonance Imaging*, vol. 19, no. 2, pp. 145–159, 2004.

[8] D. W. McRobbie, E. A. Moore, R. J. Graves, and M. R. Prince, *To BOLDly Go: fMRI, Perfusion and Diffusion*, 3rd ed. Cambridge University Press, 2017, p. 288302.

[9] T. Huisman, "Diffusion-weighted and diffusion tensor imaging of the brain, made easy," *Cancer Imaging*, vol. 10, no. 1A, p. S163, 2010.

[10] A. L. Alexander, J. E. Lee, M. Lazar, and A. S. Field, "Diffusion tensor imaging of the brain," *Neurotherapeutics*, vol. 4, no. 3, pp. 316–329, 2007.

[11] E. Schwab, H. E. Cetingül, B. Afsari, M. A. Yassa, and R. Vidal, "Rotation invariant features for HARDI," in *Information Processing in Medical Imaging*, 2013, pp. 322–330.

[12] E. Schwab, B. Afsari, and R. Vidal, "Estimation of non-negative ODFs using eigenvalue distribution of spherical functions," in *Medical Image Computing and Computer Assisted Intervention*, vol. 7511, 2012, pp. 322–330.

[13] A. Deshmane, V. Gulani, M. A. Griswold, and N. Seiberlich, "Parallel MR

imaging," *Journal of Magnetic Resonance Imaging*, vol. 36, no. 1, pp. 55–72, 2012.

[14] O. Michailovich and Y. Rathi, "On approximation of orientation distributions by means of spherical ridgelets," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2008, pp. 939–942.

[15] ——, "On approximation of orientation distributions by means of spherical ridgelets," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 461–477, 2010.

[16] D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, and M. Laval-Jeantet, "Mr imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders." *Radiology*, vol. 161, no. 2, pp. 401–407, 1986.

[17] G. Douaud, S. Jbabdi, T. E. Behrens, R. A. Menke, A. Gass, A. U. Monsch, A. Rao, B. Whitcher, G. Kindlmann, P. M. Matthews *et al.*, "Dti measures in crossing-fibre areas: increased diffusion anisotropy reveals early white matter alteration in mci and mild alzheimer's disease," *Neuroimage*, vol. 55, no. 3, pp. 880–890, 2011.

[18] F. J. Meijer, B. R. Bloem, P. Mahlknecht, K. Seppi, and B. Goraj, "Update on diffusion mri in parkinson's disease and atypical parkinsonism," *Journal of the neurological sciences*, vol. 332, no. 1, pp. 21–29, 2013.

[19] "Diffusion MRI in patients with transient ischemic attacks, author=Kidwell, Chelsea S and Alger, Jeffry R and Di Salle, Francesco and Starkman, Sidney and Villablanca, Pablo and Bentson, John and Saver, Jeffrey L, journal=Stroke, volume=30, number=6, pages=1174–1180, year=1999, publisher=Am Heart Assoc."

[20] S. Choi, D. Na, C. Chung, K. Lee, D. Na, and J. Adair, "Diffusion-weighted MRI in vascular dementia," *Neurology*, vol. 54, no. 1, pp. 83–83, 2000.

[21] M. Kubicki, R. McCarley, C.-F. Westin, H.-J. Park, S. Maier, R. Kikinis, F. A. Jolesz, and M. E. Shenton, "A review of diffusion tensor imaging studies in schizophrenia," *Journal of psychiatric research*, vol. 41, no. 1, pp. 15–30, 2007.

[22] B. G. Travers, N. Adluru, C. Ennis, D. P. Tromp, D. Destiche, S. Doran, E. D. Bigler, N. Lange, J. E. Lainhart, and A. L. Alexander, "Diffusion tensor imaging in autism spectrum disorder: a review," *Autism Research*, vol. 5, no. 5, pp. 289–313, 2012.

[23] M. Hulkower, D. Poliak, S. Rosenbaum, M. Zimmerman, and M. L. Lipton, "A decade of DTI in traumatic brain injury: 10 years and 100 articles later," *American Journal of Neuroradiology*, vol. 34, no. 11, pp. 2064–2074, 2013.

[24] B. M. Asken, S. T. DeKosky, J. R. Clugston, M. S. Jaffee, and R. M. Bauer, "Diffusion tensor imaging (DTI) findings in adult civilian, military, and sport-

related mild traumatic brain injury (mTBI): a systematic critical review," *Brain imaging and behavior*, pp. 1–28, 2017.

[25] A. R. Padhani, D.-M. Koh, and D. J. Collins, "Whole-body diffusion-weighted mr imaging in cancer: current status and research directions," *Radiology*, vol. 261, no. 3, pp. 700–718, 2011.

[26] "diffusion mri in the heart."

[27] S. Li, J. Cheng, Y. Zhang, and Z. Zhang, "Differentiation of benign and malignant lesions of the tongue by using diffusion-weighted mri at 3.0 t," *Dentomaxillofacial Radiology*, vol. 44, no. 7, p. 20140325, 2015.

[28] T. M. Schouten, M. Koini, F. de Vos, S. Seiler, J. van der Grond, A. Lechner, A. Hafkemeijer, C. Möller, R. Schmidt, M. de Rooij *et al.*, "Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease," *NeuroImage: Clinical*, vol. 11, pp. 46–51, 2016.

[29] M.-X. Huang, R. J. Theilmann, A. Robb, A. Angeles, S. Nichols, A. Drake, J. D'Andrea, M. Levy, M. Holland, T. Song *et al.*, "Integrated imaging approach with MEG and DTI to detect mild traumatic brain injury in military and civilian patients," *Journal of neurotrauma*, vol. 26, no. 8, pp. 1213–1226, 2009.

[30] W. I. Essayed, F. Zhang, P. Unadkat, G. R. Cosgrove, A. J. Golby, and

L. J. O'Donnell, "White matter tractography for neurosurgical planning: A topography-based review of the current state of the art," *NeuroImage: Clinical*, 2017.

[31] D. Tuch, T. Reese, M. Wiegell, N. Makris, J. Belliveau, and V. Wedeen, "High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity," *Magnetic Resonance in Medicine*, vol. 48, no. 4, pp. 577–582, 2002.

[32] A. Goh, C. Lenglet, P. Thompson, and R. Vidal, "Estimating orientation distribution functions with probability density constraints and spatial regularity," in *Medical Image Computing and Computer Assisted Intervention*, vol. 5761, 2009, pp. 877–885.

[33] H. E. Cetingül, M. Wright, P. Thompson, and R. Vidal, "Segmentation of high angular resolution diffusion MRI using sparse riemannian manifold clustering," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 301–317, 2014.

[34] M. Descoteaux, R. Deriche, T. Knoesche, and A. Anwander, "Deterministic and probabilistic tractography based on complex fiber orientation distributions," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 269–286, Feb. 2009.

[35] A. Einstein, *Investigations on the theory of the Brownian movement.* Dover Pubs., 1956.

[36] E. O. Stejskal and J. E. Tanner, "Spin diffusion measurements: Spin echoes

in the presence of a time-dependent field gradient," *The Journal of Chemical Physics*, vol. 42, no. 1, pp. 288–292, 1965.

[37] R. Sener, "Diffusion MRI: apparent diffusion coefficient (ADC) values in the normal brain and a classification of brain disorders based on ADC values," *Computerized medical imaging and graphics*, vol. 25, no. 4, pp. 299–326, 2001.

[38] R. Fick, D. Wassermann, M. Pizzolato, and R. Deriche, "A unifying framework for spatial and temporal diffusion in diffusion MRI," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 167–178.

[39] R. H. Fick, A. Petiet, M. Santin, A.-C. Philippe, S. Lehericy, R. Deriche, and D. Wassermann, "Multi-spherical diffusion MRI: Exploring diffusion time using signal sparsity," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 71–83.

[40] V. J. Wedeen, P. Hagmann, W.-Y. I. Tseng, T. G. Reese, and R. M. Weisskoff, "Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging," *Magnetic resonance in medicine*, vol. 54, no. 6, pp. 1377–1386, 2005.

[41] P. Basser, J. Mattiello, and D. LeBihan, "Estimation of the effective self-diffusion tensor from the NMR spin echo," *Journal of Magnetic Resonance*, vol. 103, no. 3, pp. 247–254, 1994.

[42] D. Tuch, "Q-ball imaging," *Magnetic Resonance in Medicine*, vol. 52, no. 6, pp. 1358–1372, 2004.

[43] I. Aganj, C. Lenglet, and G. Sapiro, "ODF reconstruction in Q-ball imaging with solid angle consideration," in *IEEE International Symposium on Biomedical Imaging*, 2009, pp. 1398–1401.

[44] A. Tristán-Vega and C.-F. Westin, "Probabilistic ODF estimation from reduced HARDI data with sparse regularization," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2011, pp. 182–190.

[45] O. Michailovich, Y. Rathi, and S. Dolui, "Spatially regularized compressed sensing for high angular resolution diffusion imaging," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1100–1115, 2011.

[46] A. Fuster, J. van de Sande, L. Astola, C. Poupon, and B. M. ter Haar Romeny, "Fourth-order tensor invariants in high angular resolution diffusion imaging," in *Computational Diffusion MRI Workshop (CDMRI), MICCAI*, 2011, pp. 54–63.

[47] A. Ghosh and R. Deriche, "Extracting geometrical features & peak fractional anisotropy from the ODF for white matter characterization," in *IEEE International Symposium on Biomedical Imaging*, 2011, pp. 266–271. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5872403

[48] A. Ghosh, T. Papadopoulo, and R. Deriche, "Biomarkers for HARDI: 2nd &

4th order tensor invariants," in *IEEE International Symposium on Biomedical Imaging*, 2012, pp. 26–29.

[49] Y. Gur and C. R. Johnson, "Generalized HARDI invariants by method of tensor contraction," in *IEEE International Symposium on Biomedical Imaging*, 2014, pp. 718–721.

[50] E. Schwab, M. A. Yassa, M. Weiner, and R. Vidal, "Using automatic HARDI feature selection, registration, and atlas building to characterize the neuroanatomy of beta-amyloid pathology," in *MICCAI Workshop on Computational Diffusion MRI*, 2015.

[51] H.-E. Assemlal, D. Tschumperlé, and L. Brun, "Efficient and robust computation of PDF features from diffusion MR signal," *Medical Image Analysis*, vol. 13, no. 5, pp. 715–729, 2009.

[52] J.-D. Tournier, F. Calamente, D. Gadian, and A. Connelly, "Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution," *NeuroImage*, vol. 23, no. 3, pp. 1176–1185, 2004.

[53] A. Ghosh and R. Deriche, "From second to higher order tensors in diffusion-MRI," in *Tensors in Image Processing and Computer Vision*. Springer-Verlag London, 2009, pp. 315–334.

[54] S. Wolfers, E. Schwab, and R. Vidal, "Nonnegative ODF estimation via optimal

constraint selection," in *IEEE International Symposium on Biomedical Imaging*, 2014, pp. 734–737.

[55] R. Bammer, S. L. Keeling, M. Augustin, K. P. Pruessmann, R. Wolf, R. Stollberger, H.-P. Hartung, and F. Fazekas, "Improved diffusion-weighted single-shot echo-planar imaging (EPI) in stroke using sensitivity encoding (SENSE)," *Magnetic Resonance in Medicine*, vol. 46, no. 3, pp. 548–554, 2001.

[56] R. M. Heidemann, D. A. Porter, A. Anwander, T. Feiweier, K. Heberlein, T. R. Knösche, and R. Turner, "Diffusion imaging in humans at 7T using readout-segmented EPI and GRAPPA," *Magnetic Resonance in Medicine*, vol. 64, no. 1, pp. 9–14, 2010.

[57] K. Setsompop, J. Cohen-Adad, B. Gagoski, T. Raij, A. Yendiki, B. Keil, V. J. Wedeen, and L. L. Wald, "Improving diffusion MRI using simultaneous multi-slice echo planar imaging," *Neuroimage*, vol. 63, no. 1, pp. 569–580, 2012.

[58] K. Setsompop, L. Ning, and Y. Rathi, "A combined Compressed Sensing Super-Resolution Diffusion and gSlider-SMS acquisition/reconstruction for rapid sub-millimeter whole-brain diffusion imaging," *Frontiers in Physics*, no. 9, 2016.

[59] E. Candès, "Compressive sampling," in *Pro. Int. Con. Math.*, 2006.

[60] M. Lustig, D. Donoho, and J. Pauly, "Sparse MRI: The application of com-

pressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[61] L. Ning and et al., "Sparse reconstruction challenge for diffusion MRI: Validation on a physical phantom to determine which acquisition scheme and analysis method to use?" *Medical Image Analysis*, vol. 26, no. 1, pp. 316–331, 2015.

[62] J. Cheng, D. Shen, P. J. Basser, and P. T. Yap, "Joint 6D kq space compressed sensing for accelerated high angular resolution diffusion MRI," in *Information Processing in Medical Imaging*. Springer, 2015, pp. 782–793.

[63] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[64] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[65] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[66] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky, "Sparse convolutional

neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 806–814.

[67] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[68] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[69] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[70] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[71] E. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[72] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[73] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: an algorithm for design-

ing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[74] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 391–398.

[75] M. Šorel and F. Šroubek, "Fast convolutional sparse coding using matrix inversion lemma," *Digital Signal Processing*, vol. 55, pp. 44–51, 2016.

[76] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 301–315, 2016.

[77] ——, "Boundary handling for convolutional sparse representations," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1833–1837.

[78] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 5135–5143.

[79] D. Batenkov, Y. Romano, and M. Elad, "On the global-local dichotomy in sparsity modeling," *arXiv preprint arXiv:1702.03446*, 2017.

[80] V. Papyan, Y. Romano, J. Sulam, and M. Elad, "Convolutional dictionary learning via local processing," *arXiv preprint arXiv:1705.03239*, 2017.

[81] V. Papyan, J. Sulam, and M. Elad, "Working locally thinking globally-part i: Theoretical guarantees for convolutional sparse coding," *arXiv preprint arXiv:1607.02005*, 2016.

[82] ——, "Working locally thinking globally-part ii: Stability and algorithms for convolutional sparse coding," *arXiv preprint arXiv:1607.02009*, 2016.

[83] R. Chalasani, J. C. Principe, and N. Ramakrishnan, "A fast proximal method for convolutional sparse coding," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013, pp. 1–5.

[84] E. Schwab, R. Vidal, and N. Charon, "Spatial-Angular Sparse Coding for HARDI," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2016, pp. 475–483.

[85] ——, "Joint spatial-angular sparse coding for diffusion MRI with separable dictionaries," *ArXiv*, 2017.

[86] M. Paquette, S. Merlet, G. Gilbert, R. Deriche, and M. Descoteaux, "Comparison of sampling strategies and sparsifying transforms to improve compressed sensing diffusion spectrum imaging," *Magnetic Resonance in Medicine*, vol. 73, no. 1, pp. 401–416, 2015.

[87] A. Auría, A. Daducci, J.-P. Thiran, and Y. Wiaux, "Structured sparsity for

spatially coherent fibre orientation estimation in diffusion MRI," *NeuroImage*, vol. 115, pp. 245–255, 2015.

[88] J. Cheng, D. Shen, P.-T. Yap, and P. J. Basser, "Tensorial spherical polar Fourier diffusion MRI with optimal dictionary learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 174–182.

[89] A. Daducci, E. J. Canales-Rodríguez, H. Zhang, T. B. Dyrby, D. C. Alexander, and J.-P. Thiran, "Accelerated microstructure imaging via convex optimization (AMICO) from diffusion MRI data," *NeuroImage*, vol. 105, pp. 32–44, 2015.

[90] R. Aranda, A. Ramirez-Manzanares, and M. Rivera, "Sparse and Adaptive Diffusion Dictionary (SADD) for recovering intra-voxel white matter structure," *Medical Image Analysis*, vol. 26, no. 1, pp. 243–255, 2015.

[91] I. B. Alaya, M. Jribi, F. Ghorbel, and T. Kraiem, "A Novel Geometrical Approach for a Rapid Estimation of the HARDI Signal in Diffusion MRI," in *International Conference on Image and Signal Processing*. Springer, 2016, pp. 253–261.

[92] L. Ning, K. Setsompop, O. V. Michailovich, N. Makris, M. E. Shenton, C.-F. Westin, and Y. Rathi, "A joint compressed-sensing and super-resolution approach for very high-resolution diffusion imaging," *NeuroImage*, vol. 125, pp. 386–400, 2016.

[93] S. Yin, X. You, W. Xue, B. Li, Y. Zhao, X.-Y. Jing, P. S. Wang, and Y. Tang, "A Unified Approach for Spatial and Angular Super-Resolution of Diffusion Tensor MRI," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 312–324.

[94] X. Shi, X. Ma, W. Wu, F. Huang, C. Yuan, and H. Guo, "Parallel imaging and compressed sensing combined framework for accelerating high-resolution diffusion tensor imaging using inter-image correlation," *Magnetic resonance in medicine*, vol. 73, no. 5, pp. 1775–1785, 2015.

[95] J. Sun, E. Sakhaee, A. Entezari, and B. C. Vemuri, "Leveraging EAP-Sparsity for Compressed Sensing of MS-HARDI in (k,q)-Space," in *Information Processing in Medical Imaging*. Springer, 2015, pp. 375–386.

[96] D. McClymont, I. Teh, H. J. Whittington, V. Grau, and J. E. Schneider, "Prospective acceleration of diffusion tensor imaging with compressed sensing using adaptive dictionaries," *Magnetic Resonance in Medicine*, 2015.

[97] M. Mani, M. Jacob, A. Guidon, V. Magnotta, and J. Zhong, "Acceleration of high angular and spatial resolution diffusion imaging using compressed sensing with multichannel spiral data," *Magnetic Resonance in Medicine*, vol. 73, no. 1, pp. 126–138, 2015.

[98] A. Daducci, E. J. Canales-Rodrı, M. Descoteaux, E. Garyfallidis, Y. Gur, Y.-C. Lin, M. Mani, S. Merlet, M. Paquette, A. Ramirez-Manzanares *et al.*, "Quanti-

tative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 384–399, 2014.

[99] C. Ye, "Fiber orientation estimation using nonlocal and local information," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2016, pp. 97–105.

[100] B. Yoldemir, M. Bajammal, and R. Abugharbieh, "Dictionary Based Super-Resolution for Diffusion MRI," in *MICCAI Workshop on Computational Diffusion MRI*. Springer, 2014, pp. 203–213.

[101] Y. Ouyang, Y. Chen, Y. Wu, and H. M. Zhou, "Total variation and wavelet regularization of orientation distribution functions in diffusion MRI," *Inverse Problems and Imaging*, vol. 7, no. 2, pp. 565–583, 2013.

[102] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[103] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[104] S. Merlet, J. Cheng, A. Ghosh, and R. Deriche, "Spherical polar Fourier EAP

and ODF reconstruction via compressed sensing in diffusion MRI," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2011, pp. 365–371.

[105] B. Bilgic, K. Setsompop, J. Cohen-Adad, A. Yendiki, L. L. Wald, and E. Adalsteinsson, "Accelerated diffusion spectrum imaging with compressed sensing using adaptive dictionaries," *Magnetic Resonance in Medicine*, vol. 68, no. 6, pp. 1747–1754, 2012.

[106] J. Cheng, T. Jiang, R. Deriche, D. Shen, and P.-T. Yap, "Regularized spherical polar Fourier diffusion MRI with optimal dictionary learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 639–646.

[107] A. Gramfort, C. Poupon, and M. Descoteaux, "Sparse DSI: Learning DSI structure for denoising and fast imaging," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2012, pp. 288–296.

[108] ——, "Denoising and fast diffusion imaging with physically constrained sparse dictionary learning," *Medical Image Analysis*, vol. 18, no. 1, pp. 36–49, 2014.

[109] S. Merlet, E. Caruyer, and R. Deriche, "Parametric dictionary learning for modeling EAP and ODF in diffusion MRI," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2012, pp. 10–17.

[110] S. Merlet, E. Caruyer, A. Ghosh, and R. Deriche, "A computational diffusion

MRI and parametric dictionary learning framework for modeling the diffusion signal and its features," *Medical Image Analysis*, vol. 17, no. 7, pp. 830–843, 2013.

[111] J. Sun, Y. Xie, W. Ye, J. Ho, A. Entezari, S. J. Blackband, and B. C. Vemuri, "Dictionary learning on the manifold of square root densities and application to reconstruction of diffusion propagator fields," in *International Conference on Information Processing in Medical Imaging*. Springer, 2013, pp. 619–631.

[112] S. Merlet and R. Deriche, "Compressed sensing for accelerated EAP recovery in diffusion MRI," in *MICCAI*, 2010, pp. Page–14.

[113] M. Menzel, E. Tan, K. Khare, J. Sperl, K. King, X. Tao, C. Hardy, and L. Marinelli, "Accelerated diffusion spectrum imaging in the human brain using compressed sensing," *Magnetic Resonance in Medicine*, vol. 66, no. 5, pp. 1226–1233, 2011.

[114] S. Merlet and R. Deriche, "Continuous diffusion signal, EAP and ODF estimation via compressive sensing in diffusion MRI," *Medical Image Analysis*, vol. 17, no. 5, pp. 556–572, 2013.

[115] J. Cheng, S. Merlet, E. Caruyer, A. Ghosh, T. Jiang, and R. Deriche, "Compressive sensing ensemble average propagator estimation via l1 spherical polar fourier imaging," in *MICCAI Workshop on Computational Diffusion MRI*, 2011.

[116] Y. Rathi, O. Michailovich, K. Setsompop, S. Bouix, M. Shenton, and C.-F. Westin, "Sparse multi-shell diffusion imaging," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2011, pp. 58–65.

[117] J. M. Duarte-Carvajalino, C. Lenglet, J. Xu, E. Yacoub, K. Ugurbil, S. Moeller, L. Carin, and G. Sapiro, "Estimation of the CSA-ODF using Bayesian compressed sensing of multi-shell HARDI," *Magnetic Resonance in Medicine*, vol. 72, no. 5, pp. 1471–1485, 2014.

[118] O. Michailovich and Y. Rathi, "Fast and accurate reconstruction of HARDI data using compressed sensing," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2010, pp. 607–614.

[119] J. M. Duarte-Carvajalino, C. Lenglet, K. Ugurbil, S. Moeller, L. Carin, and G. Sapiro, "A framework for multi-task Bayesian compressive sensing of DW-MRI," in *MICCAI Workshop on Computational Diffusion MRI*, 2012, pp. 1–13.

[120] B. A. Landman, J. A. Bogovic, H. Wan, F. E. Z. ElShahaby, P.-L. Bazin, and J. L. Prince, "Resolution of crossing fibers with constrained compressed sensing using diffusion tensor MRI," *NeuroImage*, vol. 59, no. 3, pp. 2175–2186, 2012.

[121] D. Kuhnt, M. H. Bauer, J. Egger, M. Richter, T. Kapur, J. Sommer, D. Merhof, and C. Nimsky, "Fiber tractography based on diffusion tensor imaging compared with high-angular-resolution diffusion imaging with compressed sensing: initial experience," *Neurosurgery*, vol. 72, no. 0 1, p. 165, 2013.

[122] D. Kuhnt, M. H. A. Bauer, J. Sommer, D. Merhof, and C. Nimsky, "Optic radiation fiber tractography in glioma patients based on high angular resolution diffusion imaging with compressed sensing compared with diffusion tensor imaging - initial experience," *PLoS ONE*, vol. 8, no. 7, pp. 1–7, 07 2013.

[123] W. Ye, B. C. Vemuri, and A. Entezari, "An over-complete dictionary based regularized reconstruction of a field of ensemble average propagators," in *IEEE International Symposium on Biomedical Imaging*. Springer, 2012, pp. 940–943.

[124] Y. Ouyang, Y. Chen, and Y. Wu, "Vectorial total variation regularisation of orientation distribution functions in diffusion weighted MRI," *International Journal of Bioinformatics Research and Applications*, vol. 10, no. 1, pp. 110–127, 2014.

[125] C. Ye and J. L. Prince, "Probabilistic tractography using Lasso bootstrap," *Medical Image Analysis*, vol. 35, pp. 544–553, 2017.

[126] Y. Rathi, O. Michailovich, F. Laun, K. Setsompop, P. E. Grant, and C.-F. Westin, "Multi-shell diffusion signal recovery from sparse measurements," *Medical Image Analysis*, vol. 18, no. 7, pp. 1143–1156, 2014.

[127] T.-C. Chao, J.-y. G. Chiou, S. E. Maier, and B. Madore, "Fast diffusion imaging with high angular resolution," *Magnetic Resonance in Medicine*, vol. 7, pp. 696–706, 2017.

[128] S. Awate and E. DiBella, "Compressed sensing HARDI via rotation-invariant concise dictionaries, flexible K-space undersampling, and multiscale spatial regularity," in *IEEE International Symposium on Biomedical Imaging*, 2013, pp. 9–12.

[129] C. F. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using kronecker bases," *Neural Computation*, vol. 25, no. 1, pp. 186–220, 2013.

[130] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 438–445.

[131] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 494–504, 2012.

[132] C. F. Caiafa and F. Pestilli, "Sparse multiway decomposition for analysis and modeling of diffusion imaging and tractography," *arXiv preprint arXiv:1505.07170*, 2015.

[133] Y. Rivenson and A. S., "Compressed imaging with a separable sensing operator," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 449–452, 2009.

[134] H. R. Goncalves, "Accelerated sparse coding with overcomplete dictionaries for image processing applications," Ph.D. dissertation, 2015.

[135] N. Qi, Y. Shi, X. Sun, and B. Yin, "TenSR: Multi-dimensional tensor sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 5916–5925.

[136] E. Candès, L. Demanet, D. Donoho, and L. Ying, "Fast discrete curvelet transforms," *Multiscale Modeling & Simulation*, vol. 5, no. 3, pp. 861–899, 2006.

[137] C. You, C.-G. Li, D. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3928–3937.

[138] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, no. 3, p. 947, 2007.

[139] F. Krahmer and R. Ward, "New and Improved Johnson-Lindenstrauss Embeddings via the Restricted Isometry Property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.

[140] Z. Tan, Y. C. Eldar, A. Beck, and A. Nehorai, "Smoothing and decomposition for analysis sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1762–1774, 2014.

[141] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[142] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," *http://arxiv.org/abs/0812.1869*, 2008.

[143] B. Haeffele and R. Vidal, "Global optimality in tensor factorization, deep learning, and beyond," *arXiv*, vol. abs/1506.07540, 2015.

[144] B. Haeffele, E. Young, and R. Vidal, "Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing," in *International Conference on Machine Learning*, 2014, pp. 2007–2015.

[145] B. D. Haeffele and R. Vidal, "Structured low-rank matrix factorization: Global optimality, algorithms, and applications," *arXiv preprint arXiv:1708.07850*, 2017.

[146] K. Gupta and S. P. Awate, "Bayesian dictionary learning and undersampled multishell HARDI reconstruction," in *Information Processing in Medical Imaging*. Springer, 2017, pp. 453–465.

[147] P. K. Pisharady, S. N. Sotiropoulos, J. M. Duarte-Carvajalino, G. Sapiro, and C. Lenglet, "Estimation of white matter fiber parameters from compressed multiresolution diffusion MRI using sparse Bayesian learning," *NeuroImage*, 2017.

[148] K. Gupta, D. Adlakha, V. Agarwal, and S. P. Awate, "Regularized dictionary learning with robust sparsity fitting for compressed sensing multishell HARDI,"

in *Computational Diffusion MRI: MICCAI Workshop*. Springer, 2016, pp. 35–48.

[149] S. St-Jean, P. Coupé, and M. Descoteaux, "Non local spatial and angular matching: Enabling higher spatial resolution diffusion MRI datasets through adaptive denoising," *Medical image analysis*, vol. 32, pp. 115–130, 2016.

[150] F. Zhang, Y. Cen, R. Zhao, and H. Wang, "Improved separable dictionary learning," in *IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 884–889.

[151] F. Zhang, Y. Cen, R. Zhao, H. Wang, Y. Cen, L. Cui, and S. Hu, "Analytic separable dictionary learning based on oblique manifold," *Neurocomputing*, vol. 236, pp. 32–38, 2017.

[152] S. Zubair and W. Wang, "Tensor dictionary learning with sparse tucker decomposition," in *IEEE 18th International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–6.

[153] G. Duan, H. Wang, Z. Liu, J. Deng, and Y.-W. Chen, "K-CPD: Learning of overcomplete dictionaries for tensor sparse coding," in *21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 493–496.

[154] F. Roemer, G. Del Galdo, and M. Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 3963–3967.

[155] A. Stevens, Y. Pu, Y. Sun, G. Spell, and L. Carin, "Tensor-dictionary learning with deep Kruskal-factor analysis," in *Artificial Intelligence and Statistics*, 2017, pp. 121–129.

[156] C. F. Dantas, M. N. da Costa, and R. da Rocha Lopes, "Learning dictionaries as a sum of Kronecker products," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 559–563, 2017.

[157] M. Bahri, Y. Panagakis, and S. Zafeiriou, "Robust Kronecker-decomposable component analysis for low-rank modeling," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[158] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "STARK: Structured dictionary learning through rank-one tensor recovery," *arXiv preprint arXiv:1711.04887*, 2017.

[159] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[160] F. Yellin, B. Haeffele, and R. Vidal, "Blood cell detection and counting in holographic lens-free imaging by convolutional sparse dictionary learning and

coding," in *IEEE International Symposium on Biomedical Imaging*, 2017, pp. 650–653.

[161] L. Bao, W. Liu, Y. Zhu, Z. Pu, and I. E. Magnin, "Sparse representation based MRI denoising with total variation," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on.* IEEE, 2008, pp. 2154–2157.

[162] T. M. Quan and W.-K. Jeong, "Compressed sensing dynamic MRI reconstruction using GPU-accelerated 3D convolutional sparse coding," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2016, pp. 484–492.

Vita

Evan Schwab received his Bachelor of Arts degree in Mathematics from Cornell University in 2010. Evan knew he wanted to continue on to pursue a PhD. but was not yet clear on what quantitative field. Evan become a Research Associate of the Megason Lab in the Department of Systems Biology at Harvard Medical School in Boston, MA from 2010-2011, conducting bioinformatic research on the zebrafish genome. There he also collaborated on an imaging project of segmenting and tracking zebrafish cells as they grew and multiplied from an embryo. It was this project that sparked Evan's interest in image processing and analysis. In the summer of 2011 Evan joined Dr. René Vidal's Vision, Dynamics and Learning Lab at The Johns Hopkins University in Baltimore, MD, where he began to pursue his Ph.D. in medical image analysis at the Center for Imaging Science through the Department of Electrical and Computer Engineering. From the start, Evan was fascinated by the impact, mathematics, and beauty of the field of diffusion magnetic resonance imaging and spent his career at Hopkins advancing this important medical imaging modality. Evan also become a student of machine learning and has applied these methods as the

backbone of his thesis work. Evan's work has been recognized in prestigious medical imaging conferences such as SIAM, ISMRM, ISBI, IPMI, and MICCAI, where he was a finalist in the best student paper award competition in 2016.

Evan's post-graduate career goals are to continue advancing medical image analysis and machine learning research in industry. In the summers of 2013 and 2014, Evan interned at Siemens Healthcare in Princeton, NJ, working on diffusion MRI and compressed sensing problems. After his Ph.D. defense in October 2017, Evan joined Philips Research in the department of Clinical Informatics, Solutions, and Services in Cambridge, MA, as a Research Scientist.